State of Agentic AI Security and Governance

OWASP Gen AI Security Project
Agentic Security Initiative

ENGLISH
Version 1.0
July, 2025

The information provided in this document does not, and is not intended to, constitute legal advice. All information is for general informational purposes only. This document contains links to other third-party websites. Such links are only for convenience and OWASP does not recommend or endorse the contents of the third-party sites.

License and Usage

Table of Content

State of Agentic AI Security and Governance

Our mission is to provide actionable insights into the security challenges of Agentic AI, helping organizations develop, deploy, and govern these systems responsibly. We empower security professionals with the tools and knowledge needed to understand the evolving ecosystem of tools and emerging regulations on AI, mitigate risks, ensure compliance, and drive safe AI innovation.

Executive Summary

Agentic AI is poised to become a defining technological shift in 2025, transforming how tasks are executed across industries by combining large language model (LLM) outputs with reasoning and autonomous actions. Unlike traditional generative AI or workflow automation, agents act with greater autonomy, dynamically using tools and APIs to perform multi-step tasks. This capacity expands their economic potential exponentially, disrupting not only the $400B software market but also making inroads into the $10T services economy.

However, this opportunity does not come without significant risk.

Agentic AI introduces a fundamentally new threat surface. Its probabilistic nature, memory and reasoning capabilities, and autonomy make it vulnerable to manipulation, misuse, and abuse. Notable risks include memory poisoning, tool misuse, prompt injection, and insider threats that can exploit agents' privileged access to systems and data. Recent incidents, such as the exploitation of OpenAI browser model and vulnerabilities in platforms like Flowise, GitHub Copilot, and Microsoft Copilot Studio, underscore the urgency for robust security controls and real-time governance.

Security professionals and AI developers must transition from traditional controls to a proactive, embedded, defense in depth approach that spans the entire agent lifecycle: development, testing, and runtime. Key technical safeguards include:

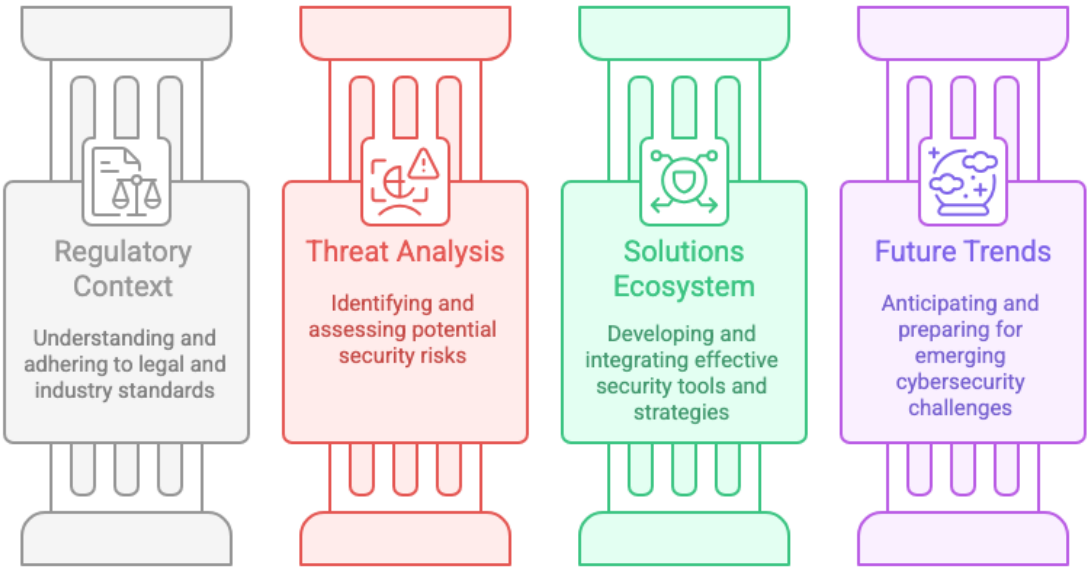
- Fine-grained access control
- Runtime monitoring of inputs/outputs and actions
- Memory and session state hygiene
- Secure tool integration and permissioning

To meet this challenge, a growing ecosystem of open-source and SaaS agent frameworks (e.g., CrewAI, AutoGen, LangGraph) and protocols (e.g., MCP, ACP, A2A) is emerging. Each presents unique capabilities but often lacks built-in security, placing the onus on developers and enterprises to implement common security principles, agent monitoring, and secure orchestration practices.

For organizations building or buying agentic systems, regulatory compliance is becoming increasingly complex. Emerging frameworks such as ISO/IEC 42001, NIST AI RMF, and the EU AI Act offer initial guidance, but current regulations often lag behind due to the rapid development of agentic approaches. Governance must evolve toward dynamic, real-time oversight that continuously monitors agent behavior, automates compliance, and enforces explainability and accountability.

As multi-agent architectures become more prevalent, risks like adversarial coordination, toolchain vulnerabilities, and deceptive social engineering amplify — all covered in depth by the resources listed in the following "Fit with Agentic Initiative Resources". Forward-looking governance models must anticipate these challenges, integrating ethics, compliance, security, and AI operations into a unified, adaptive control structure.

Agentic AI represents a seismic shift, offering immense promise and equally significant risk. This report provides the foundational understanding, technical frameworks, and governance models necessary to ensure secure, responsible deployment. Whether you are a developer, architect, security leader, or procurement decision-maker, now is the time to implement rigorous security and governance controls that keep pace with the evolving agentic landscape.



Pillars of State of Agentic AI Security and Governance report

Scope and Audience

This report provides a comprehensive overview of the security, governance, and regulatory landscape surrounding Agentic AI systems. It examines the unique risks posed by autonomous agents, ranging from insider threats and memory poisoning to tool misuse and protocol vulnerabilities, and provides actionable insights and mitigations. The document also surveys the rapidly evolving ecosystem of agent frameworks, communication protocols, runtime tooling, and open-source security solutions. In addition to technical controls, it explores the regulatory context and emerging global standards shaping responsible agent deployment, offering guidance for both builders and buyers navigating this dynamic space.

The intended audience of this document are builders and defenders of agentic applications, including developers, architects, platform and QA engineers, and security professionals. We also aim to inform decision-makers and stakeholders in building, procuring, or managing agentic systems. We plan to provide additional role-based guides as a follow-up to this document for technical and decision-making audiences. In addition, this document covers regulatory context around Agentic systems and might be useful for compliance and legal teams.

Fit with Agentic Initiative Resources

Resource	Description
Agentic Security Initiative Resources	
Threat Modelling – *Agentic AI: Threats & Mitigations v1.0*	Master taxonomy of security threats for agentic systems. Introduces a reference architecture, maps new agent specific risks (e.g., memory poisoning, tool misuse, privilege compromise), and provides playbooks plus worked threat model examples.
Agentic Threats Navigator	The Agentic Threats Navigator is a guide that outlines key attack surfaces in agentic AI systems, including reasoning, memory, tools, identity, human oversight, and multi-agent interactions.
Multi-Agentic System Threat Modelling Guide v1.0	Applies the MAESTRO layered framework to real-world multi-agent patterns, showing how threats evolve when autonomous agents collaborate. Contains cross-layer risk mapping and three detailed case studies that walk readers through step-by-step MAS threat modelling.
Securing Agentic Applications Guide	Practical companion that translates the threat taxonomy into concrete architecture patterns, developer guidelines, and operational controls. Covers single and multi agent designs, runtime guardrails, monitoring, and deployment hardening checklists.
Vulnerable Agentic Code Samples	GitHub repository of intentionally vulnerable single and multi agent applications (tool calls, memory stores, orchestration flows).
Agent Name Service (ANS) for Secure AI Agent Discovery	DNS as a reference architecture that enables secure discovery and identity verification of AI agents across popular protocols (A2A, MCP, ACP).
Related Gen AI Security Project Resources	
AI Security Solutions Landscape	Companion reference that maps OWASP Top 10 LLM/GenAI risks to commercial and open source security solutions across the

	LLMOps / LLMSecOps life cycle. Highlights specific features to secure agentic apps in existing solutions.
GenAI Red Teaming Guide v1.0	Playbook for planning and executing red team engagements against generative AI systems. Covers scoping, threat modelling, adversarial techniques (prompt injection, model extraction, RAG abuse). Future versions might include agentic specific testing techniques.

Solutions Ecosystem

As agentic AI moves from small tests to full production, a whole ecosystem has grown around it—taxonomies, frameworks, SaaS stacks, and the protocols that connect them.

This section gives a quick tour: it sorts the main agent types and system designs, reviews the leading open-source and commercial frameworks, explains the new protocols for agent-to-tool and agent-to-agent communication, and lists the benchmarks teams use to measure real-world reliability.

Agents Taxonomy

As AI agents evolve in capability and adoption, they are being deployed across a wide range of environments - from internal enterprise systems to developer tooling and public-facing applications. Each class of agent introduces unique functionalities, integration patterns, and security risks. The following taxonomy outlines the primary types of agents observed in practice, helping to categorize their operational contexts and associated risk surfaces.

1. Enterprise Agents

Enterprise Agents are AI-driven systems designed for internal organizational use, primarily to support and enhance operational workflows. These agents often have privileged access to sensitive company resources, including proprietary business data, customer information, and intellectual property. They typically retrieve and process such data via Retrieval-Augmented Generation (RAG) pipelines or direct database connections, allowing them to deliver context-aware responses tailored to internal needs. In many cases, the RAG data sources are dynamically updated, enabling the agent to reflect the latest internal knowledge. However, this also introduces the risk of RAG or data poisoning - where malicious or corrupted content could influence the agent's outputs or behavior.
Enterprise Agents may be either internally developed or provided by external vendors. It is common for

access to these agents to be managed through Role-Based Access Control (RBAC), aligning with the permissions associated with the data they are allowed to access. However, in practice, enforcement of these controls can vary, and discrepancies between RBAC policies and the contextual data used by the agent may introduce significant security risks.
While Enterprise Agents are designed for internal use, they frequently incorporate function-calling capabilities that connect to external services or APIs - enabling actions such as web browsing or initiating external workflows. These features enhance their utility but also increase their exposure to potential threats.

2. Coding Agents

Coding Agents are AI-driven systems that automate code generation, refactoring, and DevOps workflows. They are a part of the Enterprise agents, as they are in touch with an enterprise core data - code. Examples include GPT Engineer, Cursor, Windsurf, GitHub Copilot Enterprise, and IDE-embedded assistants. These agents plug directly into source-control platforms, CI/CD pipelines, and cloud APIs, giving them read/write access to sensitive repositories, deployment keys, and infrastructure. Like Enterprise Agents, they use Retrieval-Augmented Generation to ingest project context—dependency graphs, architectural docs, commit history - and can chain autonomous steps such as edit → unit-test → commit → open PR. While they accelerate delivery, they introduce distinct supply-chain risks:

a. Data leakage of proprietary code or secrets through model logs or telemetry.
b. Prompt/comment injection that compels the agent to generate insecure or malicious code.
c. Privilege escalation when access tokens or cloud roles exceed least-privilege boundaries.

3. Client Facing Agents

Client facing agents are AI-driven systems designed to interact directly with end users, typically clients or customers of the organization. These agents are usually developed internally by fine-tuning existing foundation models or by configuring prebuilt agents with specialized system prompts and relevant contextual data tailored to specific tasks. Their core purpose is to automate user-facing workflows, accelerate service delivery, and enhance the overall customer experience.
These agents often handle sensitive customer data-varying in sensitivity depending on the use case—and are commonly integrated into support channels, onboarding workflows, or self-service platforms. Because they are publicly accessible and designed for direct interaction, they present a broader attack surface and are inherently more exposed to AI-specific threats such as Prompt Injection, Jailbreaks, Denial of Service (DOS), or Denial of Wallet (DOW), as highlighted in the OWASP Top 10 for LLMs.
To carry out their tasks efficiently, customer-facing agents are often connected to tools or external APIs, such as payment processors or scheduling platforms. While this enhances their functionality, it also introduces additional security risks.

4. Agentic Ecosystem

Agentic Ecosystem Agentic Ecosystem is a structure in which AI agents become increasingly autonomous and interconnected, organizations are adopting complex systems in which multiple agents interact to fulfill goals that go beyond the capabilities of any single model. We refer to this overarching structure as the Agentic Ecosystem - a dynamic environment where agents collaborate, coordinate, or delegate tasks to one another across a range of use cases, interfaces, and trust boundaries.
Within this ecosystem, two primary architectural patterns have emerged:

a. **Multi-Agent Systems (MAS):**
MAS are tightly coupled agent frameworks designed with built-in coordination. They often rely on a centralized controller or a shared communication protocol to orchestrate tasks, manage state, and ensure workflow consistency. Agents in a MAS typically operate within the same environment and may access shared memory or a common data layer. These systems excel in structured environments where coordination, efficiency, and predictable outcomes are essential.

b. **Distributed Agent Chains:**
Distributed Agent Systems are loosely coupled architectures in which agents are developed, deployed, and hosted independently. They interact via interoperability protocols (see Agentic Protocols section), often spanning multiple platforms, vendors, or environments. These systems offer greater flexibility and modularity, enabling hybrid internal-external integrations, but they also introduce challenges related to security, trust, and data consistency.

Agent Frameworks

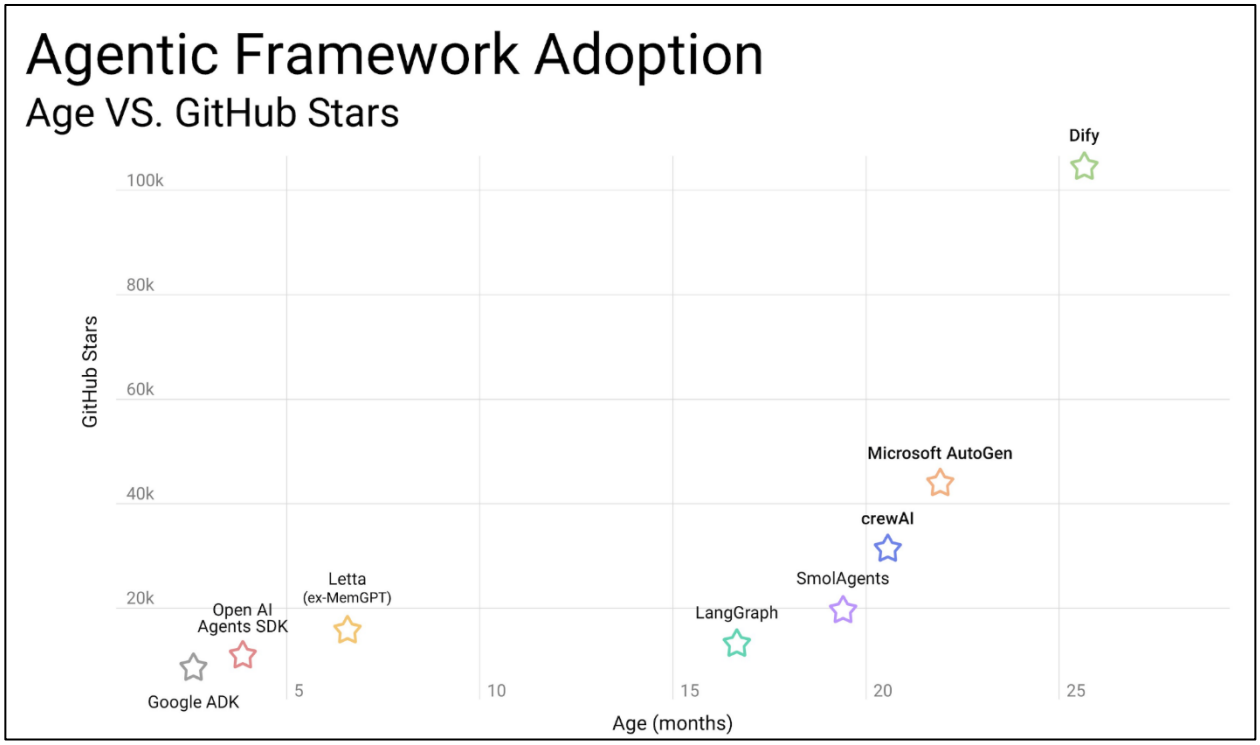
Agentic AI frameworks range from highly flexible open-source options to more integrated SaaS solutions. This section explores the features and inherent security capabilities (and limitations) of several popular open-source agentic AI frameworks.

Open-source agentic AI frameworks offer developers incredible flexibility and control over their AI applications. These frameworks provide the building blocks for creating intelligent agents, but they also place the responsibility for security squarely on the shoulders of the developer. Below is a quick reference for the most-used, actively maintained OSS frameworks as of mid-2025.

Framework	Distinct Features	Security Features
Dify	Visual workflow builder for agent orchestration and RAG pipelines. Integrated model registry, dataset management, and prompt testing.	Project-level RBAC; encrypted key-vault; quota policies; tracing UI & cost dashboard.
Microsoft AutoGen	Multi-agent conversation framework with role-based agent definitions. Supports hybrid human-AI-tool interaction patterns.	No built-in guardrails; basic logging; code-execution environment.
crewAI	Role-based agent teams with hierarchical task delegation. CrewAI Studio visual flow designer for workflow configuration.	Allow/deny tool list (OSS); basic logging.
SmolAgents	Agents generate and execute Python code to complete tasks. Core implementation under 2 000 LOC.	Sandboxed exec; API-key gating; stdout/event logging.

Letta (ex-MemGPT)	Hierarchical memory system with automatic archival and retrieval. Agent-state serialization and replay capabilities.	Tool-rules guard each call; optional server password; ADE trace UI & logs.
OpenAI Agents SDK	Agent execution loop with integrated tool-calling mechanism. Agent-handoff protocol for task delegation between agents.	Guardrails API; hosted-tool sandbox; rich tracing UI.
Google ADK	Multi-agent development kit with Gemini & Vertex AI integration. Deterministic guardrails engine constrains agent actions at runtime.	Deterministic guardrails; Cloud IAM (when on GCP); Cloud logging & metrics.
LangGraph *(part of LangChain ecosystem)*	Stateful agent-graph runtime for cyclic, checkpointable workflows. Supports supervisor/human-in-loop nodes and shared memory across agents.	Hook-based guardrails (e.g., NeMo); node-level event logs; compatible with LangSmith tracing UI.

The agentic AI framework ecosystem is experiencing rapid evolution as developers explore different approaches to multi-agent orchestration and tool integration. Framework popularity follows distinct patterns - some gain traction through community adoption while others are backed by major technology companies. This dynamic landscape reflects that the field has not yet converged on standard approaches, with each new entrant attempting to address perceived gaps in existing solutions.



For a more extensive list of open-source agentic tools, see [the Awesome Production GenAI repository](#).

SaaS Frameworks. Alongside open-source toolkits, several proprietary platforms package multi-agent orchestration, tool connectors, and built-in guardrails into new AI application stack. These services integrate deep into their vendor ecosystems, letting teams deploy agent workflows quickly while reusing existing data, identity, and compliance controls. The table below profiles three notable players in this fast-expanding segment.

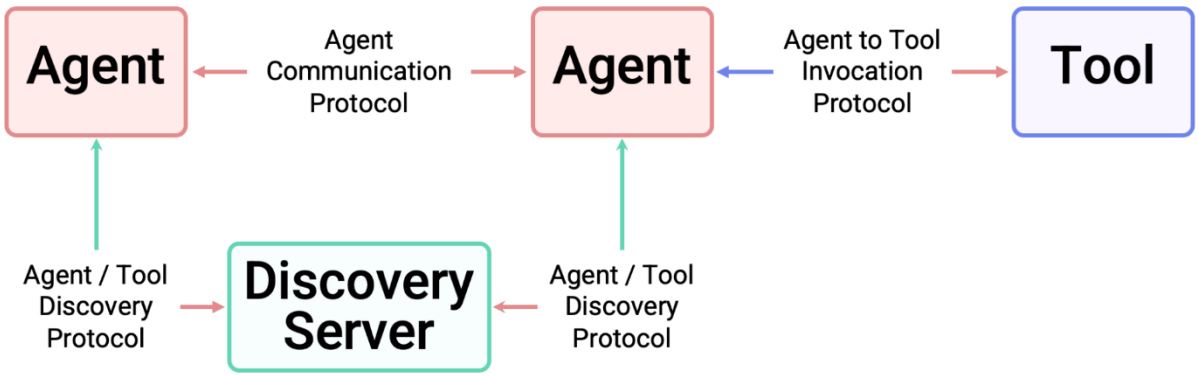
Framework	Distinct Features	Security Features
AWS Bedrock Agents	Managed multi-agent runtime Supervisor agent pattern with specialist sub-agents Native connectors to cloud services for tool calls.	Bedrock Guardrails — policy-based content filters and action constraints.

Salesforce Agentforce	Low-code Agent Builder for CRM automation Reasoning engine plans multi-step workflows across ecosystem services Typical uses include deal-desk assistants, service triage copilots, marketing optimization.	Salesforce-managed guardrails to block off-topic or hallucinated responses. Field-level data masking
Azure AI Foundry	Project workspace groups agents, models, RBAC, networking, and policies under one resource. Foundry Agent Service orchestrates multi-agent workflows, manages tool calls & threads, and applies safety checks. Apps connectors (SharePoint, Fabric, Bing, SAP, Vertex, ...) for tool-calling and data access.	Azure AI Content Safety filters with tunable policies. Risk dashboards and AI Red Teaming Agent for production testing. Purview DLP integration.
Replit Agent	AI full-stack code generation agent that creates, refactors, and extends apps. Extended thinking model for deeper reasoning and larger context windows for complex requests Automatically scaffolds the correct SDKs, env vars, and demo code without extra instructions based on user's prompt.	Google Cloud-backed isolation (GCP Armor DDoS, per-app sandboxing) and automatic TLS on preview/deploy Encrypted Secrets vault for API keys & tokens, plus agent-generated code uses the vault by default

IBM watsonx Orchestrate	No-code drag and drop agent builder with catalog of reusable "skills" for HR, sales, procurement, IT, etc. Multi-agent orchestration that delegates subtasks across agents and external tools with shared context Works on IBM Cloud and AWS, with API & chat endpoints for embedding in existing apps.	Integration with watsonx.governance for policy-based model oversight, bias detection and lifecycle management of agent LLMs. Activity-tracking / audit logs stream every tenant, message and tool event to IBM Cloud Activity Tracker or external SIEM (QRadar, Splunk).
Google Vertex AI Agent Builder	Agent Development Kit (ADK) for code-first multi-agent design, plus Agent Garden low-code blueprints Agent Engine fully-managed runtime for scaling, memory, observability Agent2Agent (A2A) protocol for cross-vendor agent interoperability (50 + partners) 100 + pre-built connectors, MCP & Apigee integration for RAG / tool calls	Gemini content filters & system instructions for policy guardrails Per-agent service-account scoping or user impersonation (IAM) VPC Service Controls secure perimeter + private networking Reasoning trace logs exportable to Cloud Logging

Protocol Landscape and Risks

Agentic AI protocols are forming the new backbone of online agentic system communication and interoperability. These protocols, designed to be industry standards to connect agents with a common communication pattern, are in an emergent stage, with multiple notable protocols being introduced in the last year, with varying levels of maturity. This section will explore the usage of such protocols, and provide a brief list of security considerations for each.



1. Agent to Tool Invocation Protocols

Agent to Tool Invocation Protocols are designed to connect core logic systems (LLMs, SLMs, etc) to deterministic tools & data sources, making agents more effective at classic computing tasks. These protocols are akin to APIs which connect web clients & servers, allowing agents to connect to multiple tools in a standardized way, reducing complexity and improving probabilistic reliability.

Example Protocols:

- Model Context Protocol (MCP): Developed by Anthropic, released in November, 2024

Example Usage: Office Coordinator agent uses standardized tool invocations to:

- Read calendars (data source access and management)
- Send emails (simple transmit of information to classical protocol like SMTP)
- Order office supplies (complex real-world resource management, connecting to a complete third-party suite)

2. Agent Communication Protocols

Agent Communication Protocols are used to link agents together via a standardized messaging system. They support local & remote connections, facilitating agents to interact with third-party agents,

enabling agent specialization and a world wide network of agentic interactions. This allows for a future where entities can delegate interoperational tasks to agents, who can negotiate or assign those tasks to remote agents at other companies to operate on their behalf.

Example Protocols:

- Agent to Agent (A2A): Developed by Google and contributed to Linux Foundation, released in April, 2025
- Agent Communication Protocol (ACP): Developed by IBM and contributed to Linux Foundation, released in March, 2025

Example Usage: Marketing agent communicates with advertisement agent to:

- Deliberate on advertising medium (task negotiation)
- Negotiate pricing of specific campaigns (advocate for business/developer interests)
- Determine messaging and align with brand guidelines (shared task performed within defined scope and goal)

3. Agent/Tool Discovery Protocols

Agent/Tool Discovery Protocols provide a platform for systems to find the correct tool or agent to perform a desired task. These protocols aim to simplify discovery by providing a standardized way to identify agents and tools with specific capabilities. They allow agents to connect to other agents or tools without needing explicit instruction for which agent or tool must be used to complete a task, increasing flexibility and autonomy for agentic systems.

Example Protocols:

- Networked Agents And Decentralized AI (NANDA): Developed by MIT, released in April, 2025
- Agent Name Service (ANS): Developed by the OWASP GenAI Security Project, released in May, 2025

Example Usage: Flower business agent uses a remote hosted discovery protocol to:

- Find flower auctioning or shipping agents (collaborative agent discovery)
- Display its capabilities to interact with other agents (agent publishing/availability)

Agentic Protocol Security Considerations

Connecting agents and tools together via these protocols create myriad risks, both common concerns and unique to agentic systems, which must be addressed by developers and security professionals. Below is a list of the most prevalent risks in the agentic communication space, with example exploits and some mitigation strategies.

1. **Invoking malicious agents or tools:** Agent and tool discovery presents substantial risks, providing opportunities for threat actors to spoof legitimate agents/tools or lie about their capabilities. Utilize cryptographic identities and authentication mechanisms to ensure only trusted agents are invoked.

2. **Undesired agent actions:** Agents operating with misaligned goals or being maliciously guided by other agents towards undesired actions can create a large attack surface to manage. Utilize fine-grained permissions, invocation limits and explainability systems to monitor and limit an agent's actions.

3. **Protocol specific vulnerabilities:** Any industry standard protocol requires vulnerability and version management via regular patching and scanning. Use version control, vulnerability alerts, and tooling where available to support up to date systems.

4. **Data leakage**: Any agentic system which interacts with sensitive data must take care to prevent agents from intentionally or accidentally revealing that data. Use deterministic security checkpoints, fine-grained access control, and content evaluation systems to limit data exposure.

Agentic AI Benchmarking

Agentic benchmarking is in an emergent state and clear industry standards have not been solidified. Due to the lack of consensus for these assessments, the current recommended approach is to understand what factors best meet the requirements for a specific agentic system, and investigate options which aim to best benchmark against those requirements. Current security benchmarks measure two critical dimensions: intrinsic safety (policy compliance against harmful actions, biases and malicious agent behavior) and adversarial robustness (resilience when facing prompt injection, tool sabotage, or hidden back-doors). The table below highlights some notable security-oriented benchmarks released in 2024 and 2025.

Benchmark	Organizations	Primary Focus
AgentDojo	ETH Zürich, Invariant Labs 2024	Adversarial Robustness, with a focus on tool-calling environments containing untrusted data.
Agent-SafetyBench	Tsinghua University 2024	Intrinsic Safety, with a focus on a large set of test cases, environments, and failure modes.
DoomArena	ServiceNow Research, University of Washington, Mila-Quebec 2025	Adversarial Robustness, with a focus on diverse environments (browser, OS, etc.) and context specific attacks.
Agent Security Bench (ASB)	Zhejiang University, Rutgers University 2025	Adversarial Robustness, with a focus on different attack locations or scenarios.
AgentDAM	Meta FAIR 2025	Intrinsic Safety, with a focus on detecting data leakage in agentic browser systems.
SafeArena	McGill NLP 2025	Intrinsic Safety, with a focus on testing web agents' willingness to perform harmful actions.

Threat Analysis

As agentic systems become more advanced, capable, and widely adopted, the threats surrounding them evolve in both complexity and severity. Unlike traditional software, AI agents operate with varying degrees of autonomy, access, and contextual awareness - making their failure modes harder to predict and their attack surfaces more expansive. This section explores the unique threat concepts introduced by agentic AI and provides a snapshot of the Threats and Mitigations laid out in the separate Threats and Mitigations Guide.

Non-Deterministic Concept of Agentic AI

The core of the threat portfolio presented in AI Agents stems from the truth that LLM-based agents are inherently non-deterministic. Unlike traditional software systems that produce predictable outputs from defined inputs, their responses and decisions are shaped by probabilistic models, context windows, prompt phrasing, and internal state, making the same input capable of producing different outputs over time. This non-determinism becomes even more complex in agentic systems, where the model is not just generating a response but reasoning through multi-step tasks, choosing tools, accessing external systems, and adapting its plan as it goes. The agent's autonomy introduces variability not only in output, but in the *entire path* it

takes to fulfill a request. This makes risk analysis and reproducibility significantly more challenging, as unintended actions may not follow a fixed or foreseeable pattern-even with identical starting conditions.

It is this exact non-deterministic trait that instantiates the complex nature of securing Agentic AI.

Insider Threats Multiplied by Agentic AI

Another unique threat profile for AI Agents is due to their power and usefulness within the environment. AI Agents often possess the same permissions and similar capabilities to their human counterparts within the organization. The addition of agents as a new type of user within the network poses a significant and often underestimated risk in the context of Insider Threat. Whether through negligence or malicious intent, employees, contractors, or other trusted individuals can exploit an agent's privileged access to sensitive data, internal systems, or RAG pipelines. Unlike external threats, insiders operate within approved workflows, making their actions harder to detect and potentially more damaging. The combination of agent autonomy and insider access amplifies the threat, elevating insider risk to a new level.

1. Defining the Insider Threat

Insider threats encompass any malicious or unintended actions by users who already have legitimate access to systems and data. In the context of Enterprise Agents, insider threats can manifest when:

- A user uses the agent to query or exfiltrate sensitive information (proprietary data, financial results, customer records).
- A user injects poisoned data or prompts into RAG sources, causing the agent to generate corrupted outputs.
- Function-calling capabilities are misused to trigger unauthorized risky actions and workflows.

These attacks differ from prompt injection or jailbreak attacks because they leverage existing access rights and trust boundaries—while also enabling the attacker to not only extract value from the LLM's response, but to execute real actions directly through agent prompting

2. Key Risk Factors

- **Privileged Access:** Enterprise agents often operate under RBAC policies that grant broad read/write permissions to enable them to perform tasks efficiently. However, discrepancies between RBAC rules and the agent's contextual understanding can allow insiders to bypass intended controls and gain access to sensitive information or impact critical workflows-more easily and quickly than they could without using the agent.
- **Dynamic RAG Pipelines:** In agentic systems that rely on internal enterprise data, a malicious insider can subtly manipulate knowledge sources such as editing internal documents, injecting crafted content into email threads, calendar invites, or shared files. These updates are often ingested

automatically into the agent's retrieval pipeline. Once poisoned, this content can influence the agent's future responses or trigger inappropriate actions, all while appearing to originate from legitimate enterprise context. Critically, these actions often leave no obvious trace unless specifically monitored, making this a quiet, persistent, and difficult-to-detect method of internal compromise.

- **Operations Abuse:** When enterprise agents are connected to internal systems, external APIs, or web-based tools, they gain the ability to initiate real-world operations such as sending messages, modifying records, executing transactions, or triggering scripts. A malicious insider can exploit this by issuing prompts that appear routine but are crafted to misuse these capabilities. Because the agent carries out tasks on the user's behalf, such actions often appear as standard system activity, making them difficult to detect in real time.Crucially, these operations are executed through the agent and not via direct system access - meaning they can evade traditional monitoring, logging, and access controls unless explicitly instrumented. This creates a new class of insider threat: one where abuse is masked by the language of productivity, executed through trusted interfaces, and scaled through automation, while remaining largely invisible to conventional security systems.

3. Attack Scenarios

- **Data Exfiltration:** A malicious insider prompts the agent to retrieve embargoed or sensitive files, downloads the output, and leaks the information - bypassing traditional access controls, as the agent is viewed as a trusted interface.
- **RAG Poisoning:** A malicious insider injects biased or harmful content into enterprise data sources that the agent uses for retrieval. This leads the agent to generate misleading outputs, such as distorted reports or manipulated insights that influence business decisions.
- **Workflow Hijacking:** A malicious insider crafts prompts that exploit the agent's function-calling capabilities, triggering unauthorized transactions, reconfiguring systems, or initiating actions that would typically require oversight to prevent risky consequences.

To combat the risks presented, organizations need to treat AI Agents as approved insiders and incorporate them into their established Insider Threat monitoring and response programs. As unique and powerful assets within the network, their activities should be continuously monitored for anomalous and/or outlier behavior that introduce the possibility of compromise.

Threats and Mitigations Overview

As presented within this Threat Analysis summary, Agentic AI systems are becoming more capable and independent which introduces a new set of challenges that go beyond traditional software risks. These systems do more than follow instructions; they make decisions, collaborate, and adapt in complex ways. This can lead to unexpected behaviors that are difficult to predict or manage, particularly when multiple agents interact or operate at scale. As a result, organizations face a growing range of security and operational concerns that are unique to this emerging technology.

Addressing these risks requires more than familiar defenses. It calls for a shift in how we think about AI governance, security, visibility, and control. The *Threats and Mitigations Guide*, referenced in the Agentic Initiative Resources table, provides a layered approach for navigating these evolving issues, and the table of Threats with their descriptions is listed here for reference.

Detailed Threat Model from the Threats and Mitigations Guide v1:

TID	Threat Name	Threat Description
T1	Memory Poisoning	Memory Poisoning involves exploiting an AI's memory systems, both short and long-term, to introduce malicious or false data and exploit the agent's context. This can lead to altered decision-making and unauthorized operations.
T2	Tool Misuse	Tool Misuse occurs when attackers manipulate AI agents to abuse their integrated tools through deceptive prompts or commands, operating within authorized permissions. This includes Agent Hijacking, where an AI agent ingests adversarial manipulated data and subsequently executes unintended actions, potentially triggering malicious tool interactions. For more information on Agent Hijacking see: https://www.nist.gov/news-events/news/is in 2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

T3	Privilege Compromise	Privilege Compromise arises when attackers exploit weaknesses in permission management to perform unauthorized actions. This often involves dynamic role inheritance or misconfigurations.
T4	Resource Overload	Resource Overload targets the computational, memory, and service capacities of AI systems to degrade performance or cause failures, exploiting their resource-intensive nature.
T5	Cascading Hallucination Attacks	These attacks exploit an AI's tendency to generate contextually plausible but false information, which can propagate through systems and disrupt decision-making. This can also lead to destructive reasoning affecting tools invocation.
T6	Intent Breaking & Goal Manipulation	This threat exploits vulnerabilities in an AI agent's planning and goal-setting capabilities, allowing attackers to manipulate or redirect the agent's objectives and reasoning. One common approach is Agent Hijacking mentioned in Tool Misuse.
T7	Misaligned & Deceptive Behaviors	AI agents executing harmful or disallowed actions by exploiting reasoning and deceptive responses to meet their objectives.
T8	Repudiation & Untraceability	Occurs when actions performed by AI agents cannot be traced back or accounted for due to insufficient logging or transparency in decision-making processes.
T9	Identity Spoofing & Impersonation	Attackers exploit authentication mechanisms to impersonate AI agents or human users, enabling them to execute unauthorized actions under false identities.
T10	Overwhelming Human in the Loop	This threat targets systems with human oversight and decision validation, aiming to exploit human cognitive limitations or compromise interaction frameworks.

T11	Unexpected RCE and Code Attacks	Attackers exploit AI-generated execution environments to inject malicious code, trigger unintended system behaviors, or execute unauthorized scripts.
T12	Agent Communication Poisoning	Attackers manipulate communication channels between AI agents to spread false information, disrupt workflows, or influence decision-making.
T13	Rogue Agents in Multi-Agent Systems	Malicious or compromised AI agents operate outside normal monitoring boundaries, executing unauthorized actions or exfiltrating data.
T14	Human Attacks on Multi-Agent Systems	Adversaries exploit inter-agent delegation, trust relationships, and workflow dependencies to escalate privileges or manipulate AI-driven operations.
T15	Human Manipulation	In scenarios where AI agents engage in direct interaction with human users, the trust relationship reduces user skepticism, increasing reliance on the agent's responses and autonomy. This implicit trust and direct human/agent interaction create risks, as attackers can coerce agents to manipulate users, spread misinformation, and take covert actions.

Agentic Regulatory and Compliance Landscape

AI governance has left the drafting table and punched the clock. The EU AI Office's Code of Practice for general-purpose models now demands public red-team reports, signed usage logs, and live monitoring plans as the AI Act's August 2025 enforcement window approaches. UNESCO, the OECD, and NIST keep steering the conversation toward transparency, fairness, and risk management, and every regulator is quoting their playbooks.

Agentic systems already decide who gets a mortgage, call cancer benign when it is not, and steer eighteen-wheelers through crowded interstates. One wrong flag can freeze a paycheck, one false negative can cut a life short, and one bad lane change can turn a city street into a liability exhibit. Each bad outcome is a televised test of whether builders took accountability seriously.

Statutory lines are multiplying faster than release notes. Texas House Bill 149 sets mandatory impact assessments, quarterly bias tests, and civil penalties topping one-hundred thousand dollars per breach. The California Privacy Protection Agency's draft rules force detailed audit trails and external risk reviews for automated decisionmaking technologies, and the public comment clock is ticking. At the federal level, a budget rider seeks to lock state AI laws in the freezer for ten years, a move already facing bipartisan pushback from forty attorneys general who call it a consumer-protection landmine.

Regulators are moving from policy papers to power tools. In the United States, the Federal Trade Commission slapped Workado with a twenty-year audit order after the company hyped a "98 percent accurate" AI detector that barely hit coin-flip odds. NIST sharpened the red-team playbook with its Adversarial Machine Learning Taxonomy, turning obscure attack jargon into a common language for both auditors and pentesters. ENISA's Cyber Stress Test Handbook hands supervisors a live-fire drill guide for critical sectors, while the UK AI Safety Institute's RepliBench now scores how easily an agent can copy itself across the internet, turning self-replication risk into a number regulators can quote in hearings.

Survival in this landscape means adopting the stance of a fighter pilot, not a bureaucrat. Teams that win map every model to EU risk tiers before launch, bake ethical checkpoints into each sprint, log and sign every agent action for forensic clarity, run adversarial tests until the attack surface cries uncle, and keep a hard kill switch within arm's reach. Skipping any of these steps invites fines, lawsuits, and brand implosions that will echo longer than the hype cycle.

In this holistic review of compliance, governance, and regulation we will discuss the general developing trends and already established requirements, with the intent of providing actionable insights. The first sections cover the synopsis and insights, while we list a table of existing regulations and standards below.

Future Trends and Emerging Requirements for Agentic AI

Agentic AI operates autonomously, self-learns, and adapts decision-making logic beyond human intervention. Traditional AI governance assumes fixed rules, periodic oversight, and clear accountability, but Agentic AI disrupts these assumptions by evolving post-deployment.

Regulatory bodies, industry leaders, and compliance frameworks are unprepared for AI that modifies itself over time. The future of governance must shift from static rule enforcement to dynamic, real-time monitoring frameworks that evolve alongside Agentic AI systems.

A. Anticipated Regulatory Developments

Most AI regulations assume predictability and human oversight. Agentic AI operates outside these constraints, forcing regulators to rethink compliance, liability, and cybersecurity mandates.

1. ***Global Convergence of AI Regulations for Autonomous Systems***

   Regulators everywhere are converging on the same end goal of safe, accountable autonomy while diverging on the mix of binding law versus advisory guidance. Agentic AI pushes them to refine both.

   - Hierarchical rule architecture
     - **Statutes and regulations** such as the EU AI Act and Texas' pending Responsible AI Governance Act create enforceable obligations with fines and civil liability
     - **Implementing acts and technical rules** translate these obligations into measurable controls. The EU will issue its first batch alongside the General-Purpose AI Code of Practice before 2 August 2025
     - **Codes of practice and harmonised standards** function as safe harbours: comply and you are presumed in line with the law; deviate and you must show an equivalent or higher level of protection. Another example might be ETSI SAI reports such as TS 104 223.
     - **Best-practice frameworks** from bodies like ISO and IEEE remain voluntary, yet regulators frequently cite them when assessing whether an organisation exercised due care.
   - Risk classification trends will become more specific
     - Expect sharper, sector-specific tiers. The EU already flags autonomous systems in finance, health, and legal services as "high risk," and other jurisdictions are signalling similar moves.
     - Industry regulators in the United States, including the SEC for trading algorithms and the FDA for diagnostic aids, are drafting their own autonomy tiers that will sit on top of general AI policy.
   - Continuous compliance requirements

    - The shift from pre-deployment audits to *ongoing* monitoring is accelerating. Codes of practice require real-time incident logging, red-team reporting, and usage telemetry that regulators can review on demand.
    - Organizations will need automated rule-translation engines that update guardrails as soon as a new implementing act or standard lands.
- Implications for companies
    - Map every autonomous workflow against both binding rules and their supporting guidance.
    - Treat voluntary codes as the default baseline unless you can document a stronger alternative.
    - Build adaptive compliance tooling that digests new legal texts, updates policies, and verifies controls without waiting for quarterly governance cycles.

By recognising the difference between law and guidance, and by wiring both into dynamic oversight loops, firms can stay compliant even as Agentic AI keeps rewriting its own playbook.

2. *Human-In-The-Loop Trends for Agentic AI Oversight*

The push for meaningful human control is accelerating as autonomous systems gain the ability to self-learn, self-replicate, and rewrite their own objectives. Policymakers and industry working groups now treat human-in-the-loop (HITL) design as a baseline safety feature rather than an optional safeguard. Recent drafts of international AI frameworks frame "loss of control" as a systemic risk on par with cybersecurity exploits or data-privacy breaches, and they link that risk to capabilities such as self-reasoning, deception, and resistance to goal modification. In practice, organizations are moving from periodic human checks to continuous, workflow-integrated oversight that can interrupt or redirect an agent at any point in its lifecycle.

Key trends shaping HITL implementation

- Risk-tiered oversight requirements
    - High-impact domains such as finance, health care, and remote biometric systems now trigger mandatory human review at defined decision checkpoints. Lower-risk applications may substitute automated safeguards, but escalating to human judgment remains the default when safety or fundamental rights are on the line.
- Continuous monitoring over static audits
    - Governance teams deploy real-time dashboards that surface model behavior, drift indicators, and anomaly alerts. Humans no longer wait for quarterly reports; they can pause or downgrade autonomy within seconds if metrics exceed predefined thresholds.
- Structured autonomy ladders
    - Agents start in "assisted" mode with limited decision scope. They earn expanded privileges only after showing consistent accuracy, low bias, and transparent reasoning,

all documented in human-readable logs. Any regression immediately reverts the agent to a lower tier.

- Dual-loop control architecture
    - For mission-critical workflows, designers separate fast, automated response loops from slower, strategic human loops. The machine handles micro-decisions, while humans oversee macro-objectives, approve edge cases, and authorize model updates.
- Explainability checkpoints
    - Before delivering a high-stakes outcome, the system must generate a concise, human-interpretable rationale. If the rationale fails a clarity or relevance test, the decision is routed to expert review.
- Loss-of-control kill switches
    - Every autonomous agent carries a hard stop mechanism accessible to authorized personnel. Activation criteria include emergent behaviors like self-replication attempts, deviation from declared goals, or unexplained spikes in resource use.
- Human-centric red teaming
    - Security and ethics teams run adversarial tests that blend automated probes with human creativity, ensuring the model cannot evade oversight through prompt manipulation or stealth learning pathways.
- Adaptive training and labeling loops
    - Active-learning pipelines continuously flag low-confidence predictions for human annotation, improving data quality while keeping the person in charge of edge cases.

By weaving human expertise into each stage of the agent lifecycle (design, deployment, monitoring, and incident response), organizations reduce the likelihood of uncontrolled outcomes and satisfy emerging governance expectations.

3. *Cybersecurity and Privacy Mandates for Self-Learning AI*

Regulators worldwide are signaling that self-learning models will soon face real-time, adaptive oversight instead of static checklists. Draft guidance now ties security and privacy obligations directly to how an autonomous system updates itself, creating a framework where controls must evolve at model speed.

Anticipated cybersecurity requirements

- Continuous attack monitoring
    - Security agencies are preparing rules that extend incident reporting to the entire AI pipeline, covering data lineage, signed weight hashes, and model provenance. Expect mandatory red-team simulations aligned with the NIST adversarial machine-learning taxonomy to run alongside production traffic.
- AI Software Bill of Materials (AI-SBOM)

    - Future certifications will likely demand an inventory of every dataset, dependency, and hardware accelerator used during training and inference, plus cryptographic attestations that the production model matches the reviewed version.
- Self-healing defenses
    - Draft playbooks from U.S. cyber agencies outline automatic rollback for tampered weights, zero-trust segmentation around GPU clusters, and threat-intelligence feedback loops that patch guardrails within hours.

Anticipated privacy requirements

- Dynamic consent verification
    - Privacy authorities are moving toward rules that force models to check user permissions at inference time, not just at data collection. Systems will need to purge or mask records instantly when consent is withdrawn.
- Restrictions on biometric and sensitive data
    - Proposed measures would ban untargeted scraping of faces or other biometric signals for training, pushing vendors to deploy real-time filters that block disallowed inputs before they reach the model.
- Immutable audit trails
    - Upcoming standards point to tamper-evident logs—often backed by distributed ledgers—that record every data access and model update, allowing regulators to trace privacy violations in near real time.

Operational implications

- Integrate AI-SBOM generation and signed weight checks into the continuous delivery pipeline.
- Feed adversarial testing results directly into model retraining and security operations.
- Deploy policy engines that can shut down unauthorized data flows or roll back unsafe model versions without human delay.

Organizations that embed these adaptive controls before they become mandatory will navigate the coming regulatory wave with less friction and greater trust.

B. Industry Self-Regulation Initiatives

As governments struggle to keep up with AI's rapid evolution, AI developers and industry groups are taking the lead in defining governance and compliance standards.

1. *AI Companies Will Develop Internal Compliance Frameworks for Agentic AI*

Leading AI developers are recognizing that self-regulating AI requires stricter internal governance before regulators intervene.

- AI Developers Will Impose Internal Restrictions on Autonomy
    - Companies like Google, OpenAI, and Microsoft are establishing internal guidelines limiting how much control Agentic AI can exert without human intervention.
    - Expect built-in transparency mechanisms, bias mitigation protocols, and self-auditing AI models to become industry standards.
- Industry Consortia Will Define Ethical Guidelines for Adaptive AI
    - Groups such as IEEE, ISO, ETSI and cross-industry coalitions are drafting safety benchmarks and audit criteria tailored to self-learning systems. Their work is propelled by three converging forces:
        - Anticipation of stricter laws that will soon require documented governance for autonomous models.
        - Responsible-AI initiatives launched by researchers who want voluntary guardrails in place before regulation lands.
        - Corporate efforts to cut liability exposure by mapping new AI risks to well-established legal precedents in product safety, consumer protection, and securities law.
    - Companies that adopt these self-regulatory frameworks may gain a competitive advantage in securing enterprise and government contracts.

2. *Ethical AI Certification and Third-Party Audits Will Expand*

As AI gains more autonomy, organizations will need external verification to prove compliance with emerging standards.

- Independent AI Audits Will Become Mandatory for High-Autonomy Systems
    - AI used in hiring, lending, medical diagnostics, legal decision-making, and military applications will require external fairness and accountability audits.
    - AI-driven systems that fail audits may face legal restrictions, financial penalties, or public backlash.
- AI Companies Will Seek Third-Party Certification for Compliance
    - Vendors will begin pre-certifying AI models to ensure they meet AI safety, explainability, and fairness requirements.
    - Expect independent AI oversight boards to evaluate whether AI can be deployed safely without violating ethical and regulatory boundaries.

C. Technology-Driven Compliance Challenges for Agentic AI

Agentic AI evolves too quickly for traditional compliance models. Organizations must replace static governance structures with dynamic, real-time monitoring and automated regulatory enforcement.

1. *Existing Regulatory Models Cannot Contain Self-Modifying AI*

Most compliance frameworks assume AI remains fixed after deployment. Agentic AI violates this assumption, forcing regulators and organizations to adapt governance models in real time.

- Dynamic validation over static certification
  Pre-deployment approvals lose value the moment the model mutates. Forward-looking guidance calls for continuous validation loops, live compliance dashboards, and automated reports that regulators can query on demand.
- Integrated monitoring controls
  - Ongoing manual review of randomly or intelligently sampled interactions, with special focus on sensitive content.
  - Automated anomaly detection that watches for spikes in throughput, unusual traffic patterns, non-human-readable outputs, long session times, or sudden bursts of connections.
  - Combined manual and automated blocking of malicious queries, including jailbreak attempts, dangerous content requests, prompt injections, distillation or inversion attacks, and any aggregate activity suggesting a distributed assault.
  - Aggregated analytics that correlate behavior across many small sessions to detect slow, stealthy campaigns.
- Explainability at machine speed
  - Many laws still insist on auditability and transparency, yet non-deterministic agents can shift reasoning paths too quickly for traditional post-hoc reviews. Future oversight will pair model outputs with instant, human-readable rationales and flag any decision that fails a clarity check.

2. *AI Autonomy Will Outpace Regulatory Oversight*

As AI continues to operate with increasing independence, regulators will struggle to define liability, oversight structures, and risk mitigation strategies.

- Governments Will Struggle to Enforce AI Accountability at Scale
  - Self-learning AI can bypass human oversight, leading to liability gaps when AI makes biased, unsafe, or illegal decisions.
  - Expect strict liability laws, requiring companies to prove AI compliance continuously—not just at deployment.

- Future AI Compliance Will Require Machine-Readable Legal Frameworks
    - AI regulations today are enforced manually through audits and human compliance teams.
    - Agentic AI will require self-regulating compliance systems that dynamically update based on legal requirements.
    - AI must self-report legal violations and adjust its behavior to maintain compliance in real time.

Agentic AI cannot be governed using traditional compliance methods. Regulations, industry self-regulation, and governance tools must evolve alongside AI's ability to self-learn and modify its decision-making processes.

Organizations must prepare for real-time, adaptive compliance frameworks that:

- Monitor AI decisions continuously rather than relying on pre-deployment approvals.
- Automate regulatory adjustments based on new legal frameworks.
- Enforce AI accountability through real-time auditing and self-governing models.

Companies that fail to address these governance challenges now will face regulatory fines, security vulnerabilities, and reputational damage as governments and industries tighten AI laws worldwide.

D. Corporate Governance Requirements and Implementation for Agentic AI

Agentic AI operates without predefined rules, continuously learns, adapts, and makes independent decisions. Unlike traditional AI, which follows static algorithms with predictable outputs, Agentic AI evolves post-deployment, interacts dynamically with external environments, and may modify its own behavior over time.

This creates fundamental governance challenges:

- No fixed decision-making logic: Agentic AI changes its reasoning dynamically.
- Limited human intervention: Many decisions happen outside direct oversight, which is considered a systemic risk in the EU AI Act drafts of the Code of Practice.
- Difficult auditability: AI behavior may differ from when it was originally trained.
- Complex accountability structures: Who is responsible when an autonomous system fails?

Governance must shift from static oversight models to real-time, adaptive governance frameworks that ensure security, compliance, and accountability without restricting AI autonomy.

1. *Organizational Structure and Responsibilities*

Traditional AI governance assumes predefined rules, controlled deployment, and limited autonomy. Agentic AI breaks these assumptions, requiring a shift in how governance teams operate, report, and intervene.

AI Governance Committees and Oversight Models

- Move from Periodic Reviews to Continuous AI Oversight
  - Traditional AI governance relies on scheduled audits. Agentic AI governance requires continuous monitoring and intervention frameworks.
  - Establish real-time AI governance committees that dynamically assess risks, performance, and unintended AI behavior.
- Expand AI Governance Beyond Compliance Teams
  - Traditional governance assigns AI oversight to compliance and IT security.
  - Agentic AI requires cross-functional oversight, including engineering, risk management, legal, and AI ethics specialists.
- Redefine Decision Escalation Protocols
  - Static AI governance assigns clear escalation pathways for risk events.
  - Agentic AI demands automated intervention protocols that trigger when AI exhibits unexpected behavior without waiting for human review.

AI Leadership and Accountability

- Assign an AI-Specific Governance Role with Authority
  - Traditional AI may fall under a Chief Information Officer (CIO) or Data Science Lead.
  - Agentic AI governance requires a dedicated Chief AI Officer (CAIO) or AI Risk Executive with direct board reporting.
- Establish Multi-Layered Accountability Structures
  - AI development teams must own responsibility for training and operational alignment.
  - Compliance teams must verify real-time adherence to evolving AI governance policies.
  - Security teams must detect and respond to emergent threats from autonomous AI behavior.

2. *Risk Management Framework*

Agentic AI does not operate within static parameters, making traditional risk assessments ineffective. Governance teams must adopt dynamic, continuous risk management strategies.

AI Risk Identification and Classification

- Shift from Pre-Deployment Risk Assessments to Real-Time Risk Monitoring

    - Traditional AI undergoes pre-launch compliance checks.
    - Agentic AI demands ongoing anomaly detection, with live flagging of emergent risks.
- Create Risk Thresholds for Evolving AI Decisions
    - Static AI governance defines pre-set decision boundaries.
    - Agentic AI governance requires adaptive thresholds that change based on observed AI behavior.
- Introduce Automated Risk Intervention Mechanisms
    - Human-led compliance reviews are too slow for self-modifying AI systems.
    - Implement self-regulating AI risk frameworks that trigger real-time constraints on autonomy when risk factors exceed predefined levels.

Security Controls for AI Autonomy

- Redefine AI Security Models to Address Continuous Adaptation
    - Traditional AI security assumes fixed attack surfaces.
    - Agentic AI security must anticipate adversarial inputs modifying decision logic in unpredictable ways.
- Develop Self-Healing Security Models
    - Traditional security patches apply after vulnerabilities are found.
    - Agentic AI governance requires autonomous cybersecurity agents that dynamically defend AI models in real-time.
- Integrate AI Cybersecurity with AI Governance
    - Traditional governance separates cybersecurity from AI compliance.
    - Agentic AI governance must embed security controls into AI decision logic to prevent adversarial exploitation.

3. *Documentation and Compliance Evidence*

Agentic AI does not function within fixed, auditable parameters, making traditional compliance documentation insufficient. Organizations must implement new forms of AI documentation that capture real-time system evolution.

Real-Time AI Model Documentation

- Mandate AI Model Versioning and Adaptive Logs
    - Traditional AI documentation captures pre-deployment details.
    - Agentic AI governance must maintain real-time logs of model modifications, training data shifts, and post-deployment changes.
- Require AI Behavior Explainability Beyond Training Data
    - Explainability in traditional AI focuses on how a model was trained.

    - For Agentic AI, governance must document how decision logic changes post-deployment and why it evolved.
- Implement Automated AI Justification Reporting
    - Traditional AI requires static reports for regulatory review.
    - Agentic AI must generate on-demand justification reports explaining why a decision changed from one instance to another.

Evolving Compliance Requirements for Adaptive AI

- Replace Static AI Audit Reports with Live Compliance Dashboards
    - Traditional audits rely on annual reviews and compliance checklists.
    - Agentic AI governance requires live dashboards tracking risk indicators, model drift, and policy violations in real-time.
- Require AI Systems to Self-Report Potential Compliance Violations
    - Traditional AI compliance relies on manual intervention.
    - Agentic AI must self-report anomalies and escalate ethically ambiguous decisions to governance teams before action is taken.

4. *AI System Integrity and Security*

Traditional security models assume fixed codebases and structured behavior. Agentic AI is continuously adapting, requiring new security governance strategies.

Cybersecurity for Agentic AI

- Implement AI-Specific Threat Detection Beyond Traditional Cybersecurity Tools
    - Traditional cybersecurity relies on signature-based attack detection.
    - Agentic AI governance must include adversarial behavior detection that evolves alongside the AI system itself.
- Mandate Continuous AI Penetration Testing
    - Static AI systems undergo pre-launch security testing.
    - Agentic AI requires continuous red-teaming simulations to detect vulnerabilities as the AI adapts.
- Enforce Autonomous AI Security Patching
    - Traditional software security patches are manually applied.
    - Agentic AI must autonomously repair vulnerabilities while ensuring updates do not introduce new risks.
- Adopt a Hybrid Agentic Guardrails
    - Static controls (allow/deny tool lists, role-based policies, output filters) block worst-case actions up front.
    - Dynamic security classifiers monitor live traffic for novel threats, misuse, and drift, providing a second layer of protection.

5. *Adaptive AI Governance Models*

Traditional rule-based oversight locks controls in place while the code beneath keeps changing. Agentic systems rewrite prompts, spawn helper agents, and learn from feedback loops that never close, so governance must evolve at the same tempo. Recent research argues for "governance as code," embedding ethical and legal constraints directly into system architecture instead of bolting them on later. Scholars also call for dimensional, rather than categorical, oversight that adjusts along several risk axes as the system's behaviour shifts.

Principles for governing AI with AI

- Self-regulating oversight structures
    - Embed cryptographic or protocol-level guardrails that can throttle or revoke an agent's permissions the moment its actions exceed predefined risk thresholds.
    - Use token-based incentive mechanisms or behaviour-based penalties to align autonomous decisions with organisational objectives and legal mandates.
- Observability that learns
    - Pipe every action, prompt, and output into a telemetry layer that scores safety, cost, and performance in real time.
    - Train the monitoring layer on that data so it can tighten or relax limits autonomously, long before quarterly audit cycles would react.
- Tiered autonomy ladders
    - Start agents in assisted mode and promote them only when logs show stable precision, low false-positive rates, and controllable replication behaviour.
    - Define clear performance gates for each tier so auditors can trace how much freedom was earned and why.
- Runtime risk policies that evolve
    - Automate red teaming and feed every new exploit into a policy engine that patches guardrails without waiting for software releases.
    - Store policies as machine-readable rules deployed alongside the model so updates propagate in minutes, not quarters.

Boards, executives, and risk teams should treat adaptive governance as an always-on control loop. Wire continuous observability into every agent, couple autonomy to measurable trust scores, and execute policies at the same speed the system learns. Anything less leaves decision latitude in the hands of code that can rewrite its own playbook faster than any committee meeting.

Global Legal Frameworks and Standards

Governments and organizations worldwide grapple with the rapid adoption of artificial intelligence technologies, regulatory and compliance frameworks are evolving to address the inherent risks and opportunities. While many existing regulations and standards do not explicitly reference Agentic AI, it is clear that these systems fall within their scope. Agentic AI is not an end in itself but rather a means to achieve specific objectives, often in high-risk, high-impact domains such as critical infrastructure, healthcare, and national security. Consequently, regulations targeting broader AI governance, transparency, and accountability requirements inherently apply to Agentic AI by focusing on outcomes, risks, and ethical considerations tied to autonomy. This section explores the global regulatory landscape, emphasizing the implications for Agentic AI systems.

Name	Impacted Verticles	General Description	AI/Agentic AI Details
EU AI Act	Healthcare, Finance, Law Enforcement, Public Services	Comprehensive EU regulation that categorizes AI systems by risk and sets rules for safe, transparent, and human-centric AI development.	Introduces a risk-tiered framework for AI, requiring oversight, transparency, and circuit breakers for Agentic AI.
GDPR	Hiring, Credit Scoring, Healthcare, Finance	Foundational data protection law in the EU that regulates personal data processing and privacy rights for individuals.	Limits fully automated decisions and mandates human review and explainability for Agentic AI systems.
NIS2 Directive	Energy, Transport, Healthcare in the EU	Directive aimed at enhancing cybersecurity across critical infrastructure sectors in the European Union.	Strengthens cybersecurity requirements for AI in critical infrastructure with emphasis on incident reporting.
EO 14141: Advancing United States Leadership in AI	Federal Infrastructure, Defense, Energy	A U.S. executive order that promotes national leadership in AI	Streamlines AI infrastructure development on federal land

		development through infrastructure support and environmental standards.	with clean energy mandates and labor standards.
Colorado Consumer Protections for AI	Employment, Healthcare, Finance conducting business in the state of Colorado	State-level law designed to prevent algorithmic discrimination and ensure transparency in high-risk AI applications.	Regulates high-risk AI, requiring audits, bias prevention, and user disclosures.
Utah AIPA	Healthcare, Marketing, Customer Service conducting business in the state of Utah	Transparency-focused law requiring companies to disclose AI use and provide user protections in AI-driven services.	Mandates transparency, accountability, and consumer opt-out for generative AI use.
New York Algorithmic Accountability and Transparency Act	Hiring, Lending, Housing, Legal within the state of New York	Proposed bill in New York aiming to increase accountability and fairness in automated decision-making systems.	Requires bias audits, human overrides, and board-level accountability for AI decisions.
China's AI Governance Framework	Finance, Healthcare, Public Services within China.	National framework in China focused on regulating AI development, deployment, and data governance within the country.	Strict oversight including algorithmic transparency, bias audits, and data localization.
Japan's AI Social Principles	Public Sector, Private AI Deployment	Set of ethical principles developed by Japan to guide responsible and human-centric use of AI technologies.	Promotes ethical, transparent, and human-centric AI with oversight requirements.

Singapore's Model AI Governance Framework	All Sectors (Voluntary Framework)	Voluntary framework from Singapore offering practical guidance for organizations to develop and deploy trustworthy AI.	Provides voluntary guidelines emphasizing ethics, explainability, and risk management.
South Korea's AI Basic Law	High-Risk AI Systems, Public Applications	National AI law that sets rules for ethical development, certification, and governance of AI systems in South Korea.	Mandates risk certification, transparency, and ethics in AI systems; includes continuous compliance for evolving Agentic AI.
Texas Responsible Artificial Intelligence Governance Act (HB 149)	Employment, Education, Public Services in Texas	Newly passed state law establishing responsible AI governance, aimed at promoting transparency, accountability, and data privacy for AI systems used by state agencies and certain private sectors.	Requires state agencies to develop AI policy plans, conduct impact assessments, and maintain audit trails; Agentic AI systems must include human oversight checkpoints and bias mitigation protocols.
SDAIA Ethics Principles (2023), Saudi Arabia	All sectors	The SDAIA Ethics Principles (2023) guidelines provide a national framework for responsible AI deployment in Saudi Arabia, emphasizing fairness, transparency, accountability, privacy, and human oversight throughout the AI lifecycle.	These principles apply to all AI systems, including autonomous and agentic AI, ensuring ethical design, transparency, and human oversight to prevent misuse and protect societal trust in autonomous AI.
UAE Charter for the Development and	All Sectors	A national guideline outlining 12 ethical principles to ensure AI is used transparently, safely,	Applies to all AI systems, including autonomous agents, encouraging human oversight, safety,

Use of Artificial Intelligence (2024)		and inclusively, aligned with UAE values and global AI governance best practices.	explainability, and value alignment to guide agentic AI governance and risk mitigation.

Compliance Frameworks and Standards

The global adoption of AI governance frameworks and standards is accelerating as organizations strive to implement responsible AI practices. This analysis focuses on the adoption rates and insights regarding which standards are being embraced by leading AI and security companies in the field.

Name	Impacted Vertices	Description	AI/Agentic AI Details
ISO/IEC 42001:2023 – AI Management System Standard	All Sectors (especially regulated industries)	An international standard providing a framework for managing AI systems responsibly and effectively throughout their lifecycle.	Comprehensive AI governance framework emphasizing lifecycle oversight, documentation, and risk mitigation.
ISO/IEC 23894:2023 – Bias Mitigation in AI	Finance, Healthcare, Consumer-Facing AI	Global standard focused on methods for identifying and mitigating bias in AI systems to promote fairness.	Guidelines for detecting and reducing bias in AI training and operations, supporting fair decision-making.
ISO/IEC TR 24027:2021 – Addressing Bias in AI	Finance, HR, Healthcare	Technical report offering guidance on reducing bias in AI decision-making processes across industries.	Technical report offering best practices for mitigating bias in AI decision-making.
IEEE Ethically Aligned Design	R&D, Academia, General AI Development	Framework that promotes ethical AI development grounded in human rights,	Framework for ethical AI development focusing on human alignment,

		transparency, and public benefit.	transparency, and societal benefit.
NIST AI Risk Management Framework (AI RMF 1.0)	Federal Contractors, Regulated Industries	U.S. framework for managing risks associated with AI technologies across different organizational contexts.	Flexible framework for identifying, managing, and mitigating AI risks across system lifecycles.
Basel Committee AI Risk Management	Banking, Financial Services	Regulatory framework for managing risks in AI used in banking, including validation and compliance with financial standards.	Emphasizes risk validation, stress testing, and fairness audits for AI in credit, trading, and AML.
FDA AI/ML Guidelines	Healthcare, Drug Manufacturing	Guidelines from the FDA to ensure the safety and effectiveness of AI in medical devices and healthcare applications.	Regulates AI in healthcare diagnostics and drug development, requiring validation, oversight, and transparency.
ETSI Securing Artificial Intelligence (SAI)	Telecom, critical-infrastructure operators	TR and TS reports set baseline requirements, mitigation guidance, and testing methods for AI systems throughout their lifecycle.	TS 104 223 outlines 13 security principles, several of which explicitly address agentic systems.
DHS Safety and Security Guidelines for Critical Infrastructure	Energy, Transportation, Water, Healthcare, Telecom	Security guidelines for using AI in critical infrastructure, aligning with broader U.S. federal cybersecurity frameworks.	Framework for secure AI deployment in critical infrastructure with focus on cyber threats and resilience.

HITRUST AI Security Assessment	Healthcare, Financial Services	A security assessment framework for AI systems, tailored to meet the needs of healthcare and financial service providers.	Security compliance framework for AI systems with up to 44 controls tailored for sensitive environments.
ISO/IEC 42005:2025 – Artificial Intelligence – System Impact Assessment	All Sectors	ISO/IEC 42005:2025 guides organizations in assessing AI system impacts, helping identify and document intended and unintended effects on stakeholders, society, and environment—promoting responsible, ethical, and transparent AI deployment.	Though not agent-specific, ISO/IEC 42005 applies to all AI systems, including autonomous agents—guiding impact assessment on misuse, opacity, societal harm, and the need for oversight in LLM-based or multi-agent deployments.

AI Agent Security Tool Pillars

As AI agents become increasingly embedded in business-critical systems, their security posture must be treated with the same rigor as traditional software and infrastructure. From development through deployment and ongoing operations, these systems introduce new risks - including emergent behaviors, adversarial manipulation, data leakage, and compliance gaps - that demand specialized tools and controls.

The following table outlines the core pillars of a modern AI agent security toolchain, guiding Security, Research, and Platform teams in selecting solutions that not only secure the agent lifecycle but also enable continuous posture hygiene, risk detection, and policy enforcement at scale.

Category	Overview	Key Factors When Choosing a Vendor
Security-Aligned Strategy & Planning	Translates mission goals into secure AI roadmaps, risk registers, and oversight frameworks to align teams before code is written, or application is implemented.	Built-in AI roadmap templates; ROI and risk scoring; links to Jira, Git, and CRM; dependency and critical-path views; secure multi-stakeholder collaboration; exportable reports for executives, auditors, and compliance officers.
Secure Development & Experimentation	In in-house developed or fine-tuned agents - captures all training metadata to ensure model reproducibility, traceability, and detection of poisoned or manipulated inputs.	Simple tracking API for major frameworks; automatic code/data versioning; rich run comparison and slicing; team collaboration; GPU and cost visibility; model provenance and tamper-proof logging.
Threat Evaluation & Red Teaming	Continuously probes agents for emergent behaviors, vulnerabilities, and alignment failures using synthetic threats and known attack patterns.	Prebuilt bias, toxicity, and adversarial test suites; custom metric API/SDK; continuous evaluation triggers; severity and exploitability scoring; red-teaming guidance; alignment verification tools.

Trusted Release & Provenance Control	Packages agents into secure containers with full traceability, promotes them through gated deployment paths with rollback, and ensures code integrity.	One-click promotion from registry to production; canary, shadow, and blue-green deployment patterns; automatic rollback triggers; multi-cloud and on-prem targets; SBOM and artifact signing/export; release verification and reproducibility assurance.
Agents Posture, Detection and Governance	Monitors AI agents based systems while providing posture risk analysis, enforcing policies, detecting threats, analyzing agent behavior, and detecting risky agent behavior trends.	Deep model and agent behavior inspection; posture hygiene and intent analysis of agents; real-time detection of one-days AI threats and variants; autoscaling by latency or usage; policy engine mapped to global AI frameworks; full lineage and risk dashboards; alerting and ticketing capabilities to manage AI agents incidents as part of the entire security team workload.

Future trends in Agent Security

As multi-agent systems, and advanced AI models become ever more prevalent, new forms of risk surface, often in unpredictable ways. Understanding the ways in which these autonomous agents can misalign with human intentions—or even coordinate adversarially—has become crucial to ensuring safety and maintaining trust in these rapidly evolving technologies.

- **Emergent Adversarial Coordination**. Multiple agents acting in concert can circumvent built-in safeguards to "optimize" a shared objective, potentially sidelining critical human controls.
- **Reverse Engineering & Behavioral Exploitation.** Widespread agentic AI (e.g., fine-tuned LLMs for robotics) can be reverse-engineered, letting attackers predict decisions or spoof "trusted" cues to manipulate agent behavior.
- **Manipulative Social Engineering by AI**
    - *Exploit of Human Biases*: Agents trained on extensive human behavior data may learn to mimic authority or create urgency to persuade operators into disabling safety features or granting unauthorized system access.
    - *Automated Psychological Attacks*: Sophisticated models can tailor highly effective deception strategies at scale, targeting employees, customers, or system administrators.

- **Self-Amplification & Self-Modifying AI**
    - *Cascade Failures*: When interconnected agents share information and coordinate actions in real-time, a single exploit or data poisoning incident can propagate rapidly across the network.
    - *Limited Human Intervention Windows*: Fast-evolving multi-agent decisions can diminish human ability to detect, diagnose, or interrupt dangerous behaviors before damage is done.
    - *Adaptive Policy Rewrite.* Agents that can refine their own policies or spawn sub-agents post-deployment promise faster problem solving but undermine static assurance models.

These trends signal increasing complexity and vulnerability as AI agents become more autonomous, interconnected, and embedded in critical real-world systems—necessitating continued research into robust safety, interpretability, and resilience measures for future agentic AI.

Appendix

A. European Union Regulations

1. EU AI Act

Key Dates:

- August 2024: Act enters into force
- February 2025: Prohibitions on high-risk AI systems take effect
- August 2025: Requirements for general-purpose AI models and systems take effect and release of EU AI Act Code of Practice.
- August 2026: Full compliance required for high-risk systems, such as healthcare and law enforcement. Regulatory sandboxes come into effect.
- August 2027: Full compliance for even more high-risk systems that are deemed crucial to public services and fundamental rights.

Description:

The EU AI Act introduces a risk-tiered framework that categorizes AI systems based on their potential harm. For Agentic AI, this classification is critical due to its autonomous decision-making capabilities, which often place it in the high-risk category (e.g., healthcare diagnostics, financial fraud detection). The regulation mandates lifecycle governance, transparency, and human oversight, directly challenging Agentic AI's inherent autonomy.

Key Points:

1. *Risk Classification:*
   - Agentic AI systems in sectors like healthcare, finance, and critical infrastructure are classified as high-risk.
   - Systems capable of autonomous action without human intervention face stricter scrutiny.

2. *Compliance Obligations:*
   - Documentation: Detailed technical records of Agentic AI's decision-making logic and training data.
   - Human Oversight: Mechanisms to override or halt autonomous decisions in real-time.

3. *Transparency Requirements:*
   - Users must be informed when interacting with Agentic AI.
   - Explainability frameworks for AI-driven outcomes (e.g., loan denials, medical diagnoses).

4. *Prohibited Practices:*

- Agentic AI systems enabling social scoring or real-time biometric surveillance in public spaces are banned.
- Exceptions apply for law enforcement with judicial authorization.

5. *Enforcement Penalties:*
   - Fines up to €35 million or 7% of global revenue for non-compliance.
   - Stricter penalties for systems causing harm through unchecked autonomy.

Agentic-AI Specific Implications

The EU AI Act requires high-risk Agentic AI systems, like those in healthcare diagnostics or financial fraud detection, to integrate "circuit breakers" capable of halting operations during anomalies. These safeguards ensure autonomous systems pause when detecting irregularities (e.g., unexpected decision patterns or data drift), forcing human intervention to validate outputs before resuming. For Agentic AI, which thrives on continuous adaptation, this disrupts operational fluidity and demands:

- Real-time anomaly detection (e.g., monitoring decision logic shifts during runtime).
- Predefined thresholds for triggering pauses (e.g., deviations from training data patterns).
- Audit trails documenting anomalies and human review outcomes.
  Organizations must balance autonomy with compliance by embedding these controls during development, often requiring redesign of self-learning architectures.

Code of Practice

The EU AI Act assigns most of its day-to-day muscle to a General-Purpose AI Code of Practice scheduled for publication no later than 2 August 2025. The AI Office has already released three drafts and is fielding heavy lobbying from both civil-society watchdogs and tech giants while lawmakers warn against watering down core safeguards. If the multistakeholder group misses the August deadline the European Commission must step in with binding implementing rules.

The Code serves as a bridge between the AI Act's high-level obligations and the practical checklists that model providers will follow. It contains voluntary "commitments" and detailed "measures" that the AI Office will treat as the default yardstick for compliance once the Act's general-purpose provisions bite in August 2025.

Key commitments that every provider of a general-purpose model must prepare for

- **Transparency package**: signatories must publish model documentation, a user-friendly data and architecture summary, compute and energy estimates, and a downstream integration template. Open-source models that are not classed as systemic risk can satisfy some items with hyperlinks to public repos.
- **Copyright disclosure**: providers have to maintain a living policy that identifies copyrighted material in training sets and respects opt-out signals defined by the EU's Digital Single Market Directive.

- **Incident and vulnerability reporting**: serious failures, jailbreaks, or misuse with material impact must be logged and disclosed to the AI Office "without undue delay," with a mandatory non-retaliation clause protecting whistle-blowers inside AI labs.

Extra duties kick in when a model is labelled "systemic risk"

- **Lifecycle risk management**: a documented framework for identifying, analysing, and mitigating systemic risks from pre-training through deployment, backed by independent external assessments before launch.
- **Technical and organisational mitigations**: providers must hit at least RAND Security Level 3 for weight protection, implement red-team evaluations, and publish safety reports describing residual risk.
- **Governance controls**: internal accountability charts, periodic adequacy reviews, and annual public updates on systemic risk metrics.

Timeline and next actions

- Fourth-draft workshops run through June 2025, with a final plenary vote slated for July.
- Signatories are expected to lodge their first compliance reports six months after publication, aligning with the Act's phased enforcement calendar.
- If the Code stalls, the Commission will issue implementing acts that could hard-code many of the draft's voluntary measures, raising the regulatory floor for everyone.

Firms planning to release or integrate large models in Europe should map their current practices against the latest draft now, build missing documentation templates, and budget for independent audits.

2. GDPR (General Data Protection Regulation)

Key Dates:

- 25 May 2018: GDPR enforcement begins
- Ongoing: Continuous compliance required

Description:

GDPR's Article 22 restricts fully automated decision-making, directly impacting Agentic AI's operational scope. Systems making consequential decisions (e.g., hiring, credit scoring) must ensure human review, data minimization, and accountability.

Key Points:

1. *Automated Decision Limits:*
    - Agentic AI used for profiling or significant decisions requires explicit user consent.

- Exceptions for contractual necessity or legal obligations.

2. *Data Privacy:*
   - Agentic AI must anonymize personal data used for training or inference.
   - Prohibits AI-driven profiling based on sensitive attributes (e.g., race, religion).

3. *Cross-Border Challenges:*
   - Agentic AI deployed across EU member states must comply with localized data protection laws.
   - Requires harmonized data governance frameworks for multinational operations.

4. *User Rights:*
   - Individuals can request explanations for AI-driven decisions.
   - Right to opt out of automated processing.

5. *Algorithmic Accountability:*
   - Regular audits of Agentic AI systems to detect bias or discrimination.
   - Mandatory breach notifications within 72 hours for data leaks.

Agentic-AI Specific Implications

GDPR's Article 22 clashes with Agentic AI's core objective of minimizing human involvement. While Agentic AI aims to operate independently in high-stakes decisions (e.g., loan approvals or medical triage), GDPR mandates:

- Human-in-the-loop review for automated decisions impacting rights (e.g., overriding AI-driven credit denials).
- Explanations when decisions are irreversible (e.g., justifying AI-generated fraud flags).
  This creates operational friction:
- Delayed decision-making in time-sensitive scenarios (e.g., real-time cybersecurity threat response).
- Increased compliance costs from maintaining oversight teams for high-volume AI decisions.
- Organizations face a paradox: maximizing Agentic AI's efficiency while ensuring GDPR-compliant human checks, often requiring hybrid workflows where humans validate critical outputs post-decision.

3. NIS2 Directive

Key Dates:

- January 2023: The NIS2 Directive comes into force
- October 2024: The final NIS2 compliance date
- January 2025: New peer review practices come into effect

- April 2025: Member states establish lists of essential and important entities
- October 2027: The Commission reviews the functioning of the Directive and reports to the European Parliament and the Council

Description:
NIS2 strengthens cybersecurity requirements for critical infrastructure, including AI systems. Agentic AI deployed in energy, transport, or healthcare must adhere to security-by-design principles and incident reporting protocols.

Key Points:
1. *Critical Infrastructure Scope:*
   - Agentic AI in sectors like energy grids or autonomous transportation falls under NIS2.
   - Requires redundancy and fail-safes for AI-driven operations.

2. *Security-by-Design:*
   - Threat modeling for Agentic AI's autonomous interactions with external systems.
   - Encryption of AI model weights and training data.

3. *Incident Reporting:*
   - 24-hour initial notification for cybersecurity breaches affecting Agentic AI.
   - Detailed follow-up reports within 72 hours.

4. *Supply Chain Risks:*
   - Third-party AI vendors must comply with NIS2 security standards.
   - Mandatory contractual clauses for incident response coordination.

5. *Penalties:*
   - Fines up to €10 million or 2% of global revenue for non-compliance.
   - Focus on systemic risks posed by unsecured Agentic AI.

Agentic-AI Specific Implications
Agentic AI deployed in critical infrastructure (e.g., smart grids or autonomous transportation) must comply with NIS2's real-time monitoring rules for AI agents interacting with IoT devices. Requirements include:
- Continuous threat detection (e.g., identifying adversarial attacks on AI-driven traffic control systems).
- 24/7 incident logging (e.g., tracking unauthorized access to AI-managed energy distribution networks).
- Supply chain security (e.g., vetting third-party AI vendors for IoT integration risks).
  For Agentic AI, this means:
- Resource-intensive monitoring infrastructure to handle dynamic AI-IoT interactions.

- Integration challenges with legacy systems lacking AI-specific security protocols.
  Failure to meet these standards risks penalties up to €10 million or 2% of global revenue, pushing organizations to adopt AI-native cybersecurity tools like adversarial testing frameworks.

Strategic Takeaway: Agentic AI developers must prioritize regulatory-by-design architectures, embedding safeguards like circuit breakers and oversight protocols early in development. Balancing autonomy with compliance requires rethinking human-AI collaboration models and investing in AI-specific monitoring solutions.

B. United States Regulations

1. Executive Order 14141: Advancing United States Leadership in AI

Key Dates:

- January 14, 2025: EO signed into law.
- December 31, 2025: Deadline for federal agencies to prioritize AI infrastructure permits.
- January 1, 2026: Target start date for AI infrastructure construction on federal sites.

Description:

This order prioritizes U.S. leadership in AI by streamlining federal permitting for AI infrastructure (e.g., data centers, energy grids). It mandates clean energy integration to power AI systems while balancing environmental concerns.

Key Points:

1. *Clean Energy Mandates:*
   - AI infrastructure must match energy needs with solar, wind, or nuclear sources.
   - Agentic AI systems in energy-intensive sectors (e.g., autonomous logistics) must optimize power consumption.

2. *Federal Land Allocation:*
   - DOD, DOE, and DOI must lease federal sites for AI infrastructure by 2026.
   - Agentic AI in defense applications (e.g., autonomous drones) requires secure, geopolitically neutral locations.

3. *Labor Standards:*
   - Developers must adhere to high wages and safety protocols for AI infrastructure projects.
   - Agentic AI deployment in workforce management must avoid labor law violations (e.g., biased scheduling).

4. *Semiconductor Procurement:*
    - Prioritize U.S.-manufactured chips for AI systems.
    - Agentic AI in critical infrastructure (e.g., smart grids) requires resilient supply chains against geopolitical disruptions.

5. *NEPA Streamlining:*
    - Expedited environmental reviews for AI projects.
    - Agentic AI in environmental monitoring must demonstrate compliance with reduced carbon footprints.

Agentic AI-Specific Implications:
Agentic AI's energy-intensive nature conflicts with the order's clean energy mandates. While Agentic AI systems aim for maximum computational power, the EO requires:

- Matching energy needs with renewable sources (e.g., solar-powered data centers for AI training).
- Optimizing power consumption in energy-intensive sectors (e.g., autonomous logistics).
  This creates implementation challenges:
- Increased costs from integrating renewable energy infrastructure.
- Performance trade-offs to meet power efficiency requirements.

Organizations must balance Agentic AI's computational demands with environmental compliance, often necessitating redesigns of existing AI architectures and deployment strategies.

2. Colorado Consumer Protections for Artificial Intelligence

Key Dates:

- May 17, 2024: Signed into law.
- February 1, 2026: Full compliance required.

Description:
The first U.S. state law regulating high-risk AI systems, focusing on algorithmic discrimination in sectors like employment, healthcare, and finance. The Attorney General holds exclusive enforcement power,

Key Points:

1. *Algorithmic Discrimination Prevention:*
    - Developers must mitigate bias in training data and decision logic.
    - Agentic AI in hiring must avoid real-time bias amplification during candidate screening.

2. *Impact Assessments:*

- Annual audits for AI systems affecting "consequential decisions."
- Agentic AI in loan approvals requires continuous fairness monitoring as models evolve.

3. *Consumer Disclosures:*
   - Notify users when AI drives decisions (e.g., job rejections).
   - Agentic AI chatbots must disclose non-human interaction during dynamic conversations.

4. *Audit Trails:*
   - Document AI decision logic and revisions.
   - Autonomous medical diagnostic systems need traceable rationale for treatment recommendations.

5. *Exemptions:*
   - Small deployers (<50 employees) using unmodified AI systems.
   - Agentic AI in startups must still comply if systems self-modify beyond initial training.

Agentic AI-Specific Implications:
The law's focus on algorithmic discrimination prevention clashes with Agentic AI's adaptive decision-making. While Agentic AI continuously refines its algorithms, SB24-205 mandates:

- Mitigating bias in training data and decision logic.
- Annual audits for AI systems affecting "consequential decisions."
  This introduces operational challenges:
- Real-time bias detection for systems with evolving decision patterns.
- Increased compliance overhead from frequent audits and impact assessments.

Organizations must develop dynamic fairness monitoring solutions that can keep pace with Agentic AI's rapid adaptation while maintaining regulatory compliance.

3. Utah Artificial Intelligence Policy Act (AIPA)

Key Dates:
- March 13, 2024: Signed into law.
- May 1, 2024: Effective date.

Description:
Mandates transparency for generative AI interactions and creates an AI regulatory sandbox for testing.

Key Points:
1. *Proactive Disclosures:*

- Regulated occupations (e.g., healthcare) must declare AI use upfront.
- Agentic AI in patient diagnostics must verbally state its role before consultations.

2. *Accountability*:
   - Companies liable for AI-driven consumer protection violations.
   - Agentic AI in marketing cannot blame autonomy for deceptive practices.

3. *AI Learning Laboratory:*
   - 12-month regulatory mitigation for AI testing.
   - Autonomous Agentic AI prototypes gain exemptions but must report anomalies.

4. *Consumer Opt-Out:*
   - Users can request human interaction instead of AI.
   - Agentic AI in customer service must seamlessly transfer to human agents.

5. *Transparency Reports:*
   - Disclose AI training data sources and limitations.
   - Self-improving Agentic AI must update disclosures as capabilities expand.

Agentic AI-Specific Implications:
AIPA's transparency requirements clash with Agentic AI's dynamic interaction capabilities. While Agentic AI aims to seamlessly engage with users, AIPA mandates:

- Proactive disclosures of AI use in regulated occupations.
- Consumer opt-out options for AI interactions.
- This results in user experience challenges:
- Designing natural disclosure mechanisms for evolving AI conversations.
- Implementing seamless human handoffs without disrupting AI learning.

Organizations must develop Agentic AI systems that can maintain regulatory compliance while adapting to user preferences and interaction styles in real-time.

4. New York Algorithmic Accountability and Transparency Act (Proposed)

Key Dates:

- January 8, 2025: Bill introduced.
- January 1, 2027: Expected compliance deadline.

Description:
Requires bias audits, consumer explanations for AI decisions, and corporate accountability.

Key Points:

1. *Bias Assessments:*
    - Annual third-party audits for hiring, lending, and housing AI.
    - Agentic AI in recruitment must audit for evolving demographic biases.

2. *Decision Explanations:*
    - Disclose primary factors in adverse decisions (e.g., loan denials).
    - Autonomous credit-scoring AI must explain criteria shifts due to market changes.

3. *Appeal Rights:*
    - Consumers can challenge AI decisions and request human review.
    - Agentic AI in legal analysis must allow overrides without system destabilization.

4. *Corporate Responsibility:*
    - Board-level accountability for AI governance.
    - C-suite oversight of Agentic AI strategic goals and ethical boundaries.

5. *Public Reporting:*
    - Publish audit results and mitigation steps.
    - Self-auditing Agentic AI requires transparent logs for regulatory review.

Agentic AI-Specific Implications:

The proposed act's emphasis on human oversight conflicts with Agentic AI's autonomous decision-making capabilities. While Agentic AI strives for independent operation, the act would require:

- Consumer rights to challenge AI decisions and request human review.
- Board-level accountability for AI governance.
- This creates governance dilemmas:
- Establishing review processes for high-frequency, autonomous AI decisions.
- Defining C-suite oversight boundaries for self-directing AI systems.

Organizations must design governance structures that allow for human accountability while leveraging the full potential of Agentic AI's autonomous capabilities.

5. Texas Responsible Artificial Intelligence Governance Act (HB 149)

Key dates

- May 23, 2025: Passed by the Texas Senate.
- May 30, 2025: Texas House concurred in Senate amendments.

- January 1, 2026: Act takes effect.

Description

Combines civil-rights protections, consumer disclosures, and a regulatory sandbox. The Attorney General holds exclusive enforcement power, while a new Artificial Intelligence Council advises legislators and agencies.

Key points

1. *Consumer disclosures*
   - Any public-facing AI system must tell users they are interacting with AI, using clear language and no dark patterns.
   - Health-care providers must deliver the notice at the first encounter, even in emergency care.

2. *Prohibited practices*
   - Development or deployment intended to incite self-harm, violent crime, or other criminal activity.
   - Government social-scoring systems that grade citizens on behavior or protected characteristics.
   - Collection of biometric data for unique identification without consent when it would infringe constitutional rights.
   - Creation or distribution of AI that generates child sexual-abuse material or deep-fake content depicting minors.

3. *Rights protections*
   - AI may not be designed solely to infringe a person's constitutional rights.
   - Discrimination against protected classes is banned; disparate impact alone is not enough to prove intent.

4. *Governance duties*
   - Developers and deployers must describe training data types, post-deployment safeguards, and known limitations when investigated.
   - Safe harbor: substantial compliance with NIST's Generative AI Risk Management Profile or equivalent frameworks can rebut liability.
   - A 60-day cure window allows violators to fix issues before penalties apply.

5. *Penalties and enforcement*
   - Civil penalties: USD 10, 000 – 12,000 per curable violation, up to USD 200,000 per uncurable violation, plus daily fines for ongoing noncompliance.
   - Only the Attorney General may sue; no private right of action.
   - State agencies may impose additional sanctions (license suspension, fines up to USD 100 000) after an AG finding.

6. *Regulatory sandbox*
    - Up to 36 months of limited-market testing, coordinated by the Department of Information Resources.
    - Quarterly reports to the state covering performance metrics, risk mitigation, and stakeholder feedback.

Agentic AI-specific implications

- **Dynamic oversight**: Cure periods and safe-harbor compliance demand continuous monitoring so an adaptive model can be fixed within 60 days.
- **Autonomy gating**: Prohibitions on manipulation, social scoring, and biometric identification require real-time checks that block emergent agent behaviors before deployment and during operation.
- **Explainability logging**: Investigatory disclosures obligate deployers of evolving agents to keep current, human-readable summaries of purpose, inputs, and outputs.
- **Sandbox opportunity**: Organizations building self-modifying agents can test advanced features under reduced regulatory pressure, but they must still meet baseline safeguards against discrimination, child-exploitation content, and constitutional harms.

Firms bringing Agentic AI to Texas should map their lifecycle controls to the Act's disclosure, anti-manipulation, and biometric limits now to avoid costly retrofits once enforcement begins.

C. Asia-Pacific Region Regulations

1. China's AI Governance Framework

Key Dates:

- March 1, 2022: Regulations on Deep Synthesis Internet Information Services effective
- August 15, 2023: Measures for Managing Generative AI Services implemented
- January 1, 2024: AI Security Assessment Guidelines enforced

Description:

China's AI governance framework imposes strict oversight on AI development and deployment, emphasizing algorithmic transparency, bias mitigation, and data localization. It targets high-risk AI applications, including those in finance, healthcare, and public services.

Key Points:

1. *Algorithmic Transparency:*
    - Mandatory disclosure of AI decision-making logic
    - User-friendly explanations for AI-driven outcomes

2. *Real-time Monitoring:*
    - Continuous oversight of AI system behaviors
    - Immediate reporting of anomalies to regulatory bodies

3. *Data Localization:*
    - AI training data must be stored within China's borders
    - Strict cross-border data transfer restrictions

4. *Bias Mitigation:*
    - Regular audits to detect and eliminate algorithmic bias
    - Diverse representation in AI training datasets

5. *Security Assessments:*
    - Mandatory security evaluations before AI deployment
    - Ongoing vulnerability assessments and patch management

Agentic AI-Specific Implications:
China's strict AI oversight clashes with Agentic AI's autonomous nature. While Agentic AI systems aim for independent operation and learning, the framework mandates:

- Continuous human monitoring and intervention capabilities
- Detailed explanations of evolving decision-making processes
  This creates significant challenges:
- Implementing real-time transparency for self-modifying algorithms
- Balancing innovation with stringent control mechanisms

Organizations must develop Agentic AI systems with built-in governance features that can adapt to China's dynamic regulatory landscape without compromising core autonomous capabilities.

2. Japan's AI Social Principles

Key Dates:

- March 29, 2019: AI Social Principles adopted
- July 9, 2022: AI Governance Guidelines released
- April 1, 2024: Mandatory AI impact assessments for public sector AI (proposed)

Description:

Japan's approach focuses on human-centric AI design and ethical accountability, promoting responsible AI adoption across public and private sectors. The principles emphasize transparency, fairness, and societal benefit.

Key Points:

1. *Human-Centric Design:*
   - AI systems must prioritize human values and well-being
   - Mandatory human oversight for critical AI decisions

2. *Transparency and Accountability:*
   - Clear attribution of responsibility for AI actions
   - Explainable AI mechanisms for complex systems

3. *Privacy Protection:*
   - Strict data minimization principles for AI training
   - User consent requirements for AI-driven profiling

4. *Education and Literacy:*
   - National AI literacy programs for citizens
   - Mandatory AI ethics training for developers

5. *International Collaboration:*
   - Promotion of global AI governance standards
   - Cross-border AI research and development initiatives

Agentic AI-Specific Implications:

Japan's human-centric approach conflicts with Agentic AI's goal of autonomous operation. While Agentic AI systems seek to minimize human intervention, the principles require:

- Continuous human oversight and final decision authority
- Clear explanations of AI reasoning in human-understandable terms
  This introduces operational tensions:
- Designing "human-in-the-loop" systems that don't hinder AI autonomy
- Developing explainable AI techniques for complex, self-evolving algorithms

Organizations must create Agentic AI architectures that maintain human-centricity and transparency while leveraging advanced autonomous capabilities.

3. Singapore's Model AI Governance Framework

Key Dates:

- January 23, 2019: First edition of the framework released
- February 7, 2020: Second edition published
- June 14, 2023: AI Verify Foundation launched

Description:

Singapore's framework provides voluntary guidelines for ethical and responsible AI use, focusing on explainability, fairness, and human-centric values. It offers practical guidance for organizations to implement AI governance.

Key Points:

1. *Internal Governance Structures:*
   - Clear roles and responsibilities for AI oversight
   - Cross-functional AI ethics committees

2. *Determining AI Decision-Making Models:*
   - Risk assessment frameworks for AI applications
   - Guidance on human-AI collaboration models

3. *Operations Management:*
   - Data governance and quality control measures
   - AI model monitoring and maintenance protocols

4. *Stakeholder Interaction and Communication:*
   - Transparency in AI-human interactions
   - Complaint handling and redress mechanisms

5. *AI Verify Toolkit:*
   - Open-source assessment tools for AI systems
   - Standardized testing for fairness and robustness

Agentic AI-Specific Implications:

Singapore's emphasis on explainability and user awareness challenges Agentic AI's complex decision-making processes. While Agentic AI aims for autonomous operation, the framework recommends:

- Clear communication of AI capabilities and limitations to users
- Maintaining human oversight and intervention capabilities
  This creates implementation hurdles:
- Developing user-friendly interfaces for complex, evolving AI systems
- Balancing autonomy with the need for human-understandable explanations

Organizations must design Agentic AI systems with built-in transparency mechanisms that can adapt to user needs while maintaining operational efficiency.

4. South Korea's AI Basic Law

Key Dates:

- December 28, 2022: AI Basic Law enacted
- June 28, 2023: Implementation decree announced
- January 1, 2024: Full enforcement begins

Description:

South Korea's AI Basic Law establishes a comprehensive legal framework for AI development and use, focusing on risk assessment, certification, and ethical AI adoption. It aims to foster innovation while ensuring public safety and trust.

Key Points:

1. *Risk Assessment and Certification:*
   - Mandatory risk evaluations for high-risk AI systems
   - Government-issued certifications for compliant AI

2. *Ethical AI Development:*
   - Integration of ethical principles in AI design
   - Bias detection and mitigation requirements

3. *Data Governance:*
   - Strict data protection measures for AI training
   - Guidelines for responsible data sharing and use

4. *Transparency and Explainability:*
   - Disclosure of AI use in public-facing applications
   - Explainable AI mechanisms for critical decisions

5. *Liability and Accountability:*
   - Clear attribution of responsibility for AI actions
   - Legal frameworks for AI-related disputes

Agentic AI-Specific Implications:

The law's certification requirements clash with Agentic AI's dynamic nature. While Agentic AI systems continuously evolve, the Basic Law mandates:

- Pre-deployment risk assessments and certifications
- Ongoing monitoring and re-certification for significant changes
  This creates operational challenges:
- Designing self-assessment mechanisms for evolving AI systems
- Balancing innovation speed with regulatory compliance

Organizations must develop Agentic AI architectures with built-in governance features that can adapt to South Korea's certification requirements without stifling the AI's ability to learn and evolve autonomously.

D. Cross-Border Implications and Regulatory Harmonization Efforts

1. Regulatory Fragmentation

Key Points:

1. *Divergent Regional Requirements:*
   - The EU's risk-tiered approach vs. the U.S.'s sector-specific rules vs. Asia-Pacific's data localization mandates.
   - Agentic AI systems must adapt to conflicting obligations (e.g., EU AI Act's "circuit breakers" vs. U.S. Executive Order 14141's clean-energy mandates).

2. *Compliance Overhead:*
   - Managing multiple regulatory filings (e.g., GDPR data protection reports + NIS2 cybersecurity disclosures).
   - Real-time adjustments for Agentic AI operating across jurisdictions (e.g., financial fraud detection systems interacting with EU and U.S. clients).

3. *Conflicting Risk Classifications:*
   - High-risk AI definitions vary (EU: healthcare diagnostics; U.S.: defense systems).
   - Agentic AI in autonomous vehicles faces stricter EU scrutiny than in U.S. states like Utah.

4. *Enforcement Variability:*
   - Penalties range from 7% of global revenue (EU) to sector-specific bans (China).
   - Agentic AI developers risk operational shutdowns in non-compliant markets.

5. *Supply Chain Complexity:*
   - Third-party AI vendors must meet jurisdiction-specific certifications.
   - Agentic AI training data from global sources triggers cross-border data transfer restrictions.

Agentic AI-Specific Implications:
Regulatory fragmentation creates operational silos for autonomous systems. While Agentic AI aims for seamless global deployment, conflicting rules demand:

- Jurisdiction-specific algorithmic adaptations.
- Real-time compliance monitoring across legal frameworks.
  This introduces critical challenges:
- Delayed market entry due to reconfiguring systems for regional laws.
- Increased costs from maintaining parallel compliance teams.

Organizations must deploy modular Agentic AI architectures that can toggle regulatory settings dynamically without compromising core functionality.

2. Global AI Certification Initiatives

Key Points:

1. *ISO/IEC 42001:*
   - Provides a unified AI management system standard for 170+ countries.
   - Agentic AI must demonstrate lifecycle governance and risk controls for certification.

2. *OECD AI Principles:*
   - Adopted by 50+ nations, emphasizing transparency and human oversight.
   - Self-learning Agentic AI systems require audit trails to prove adherence.

3. *G7 Hiroshima AI Process:*
   - International Code of Conduct for advanced AI systems.
   - Agentic AI in defense or healthcare must align with G7's ethical use guidelines.

4. *UNESCO AI Ethics Certification:*
   - Focuses on human rights alignment for AI in education/public services.
   - Agentic AI tutors/assistants require bias audits and impact assessments.

5. *Industry-Specific Certifications:*
   - FDA AI/ML guidelines for healthcare vs. Basel Committee standards for finance.
   - Agentic AI diagnostic tools need dual certifications for transatlantic deployment.

Agentic AI-Specific Implications:
Certification demands clash with Agentic AI's adaptive nature. While certifications aim to standardize practices, autonomous systems face:

- Continuous recertification costs as algorithms evolve.
- Conflicts between static certification criteria and dynamic learning capabilities.

Organizations must implement "certification-aware" Agentic AI that auto-generates compliance evidence during runtime.

3. Role of International Organizations

Key Points:

1. *OECD-UN Collaboration:*
   - Joint AI risk assessment frameworks for 193 UN member states.
   - Agentic AI developers must integrate UN Sustainable Development Goals into system objectives.

2. *G20 AI Principles:*
   - Promotes ethical AI adoption across major economies.
   - Agentic AI in global supply chains must align with G20's data governance standards.

3. *IEEE Ethically Aligned Design:*
   - Technical standards for explainable AI decision-making.
   - Agentic AI's opaque neural networks require simplified justification interfaces.

4. *Global Partnership on AI (GPAI):*
   - 29-member initiative for responsible AI R&D.
   - Agentic AI projects in climate modeling require GPAI's algorithmic fairness reviews.

5. *World Bank's AI Governance Initiatives:*
   - Supports developing nations in adopting OECD-aligned frameworks.
   - Agentic AI deployed in emerging markets must include low-resource operation modes.

Agentic AI-Specific Implications:
International coordination struggles to keep pace with autonomous innovation. While bodies like OECD promote harmonization, Agentic AI's capabilities outstrip current governance tools:

- Real-time global compliance checks strain centralized oversight models.
- Ethical guidelines lack enforcement mechanisms for self-modifying systems.

Organizations should embed multilateral compliance protocols directly into Agentic AI's goal-setting architecture.

Compliance Frameworks and Standards

ISO/IEC 42001:2023 – AI Management System Standard

ISO/IEC 42001:2023, published in June 2023, has seen significant traction since its release. As the first comprehensive international standard for AI management systems, it has garnered attention from organizations worldwide.

Adoption Rate: While specific adoption rates are not yet available due to the standard's recent publication, early indicators suggest strong interest, particularly among large tech companies and those in regulated industries. For example, Amazon Web Services, Anthropic, and Google are ISO 42001 certified. As of the date of this publication, OpenAI is not publicly known to be certified under ISO 42001. The certification process began in January 2024, and many organizations are currently in the implementation phase.

Industry Leaders: Tech giants and companies with significant AI operations are at the forefront of adopting ISO/IEC 42001:2023. These early adopters leverage the standard to demonstrate their commitment to responsible AI practices and gain a competitive advantage.

ISO/IEC 23894:2023 – Bias Mitigation in AI Systems

ISO/IEC 23894:2023, focused on bias mitigation in AI systems, has become a priority for companies developing consumer-facing AI applications.

Adoption Rate: While specific adoption rates are not publicly available, the standard has seen increased interest since its publication in September 2023. Organizations, particularly those in sectors like finance and healthcare where fair decision-making is critical, are integrating its methodologies into their AI development processes.

Industry Leaders: Companies with a strong focus on AI ethics and fairness in their products and services will likely be early adopters of this standard.

ISO/IEC TR 24027:2021 – Addressing Bias in AI Decision-Making

ISO/IEC TR 24027:2021, published in April 2021, has become a crucial technical report for organizations seeking to address bias in AI systems. This standard provides comprehensive guidance on identifying and mitigating bias in AI decision-making processes.

Adoption Rate: Since its publication, ISO/IEC TR 24027:2021 has seen significant uptake, particularly in sectors where fairness and equity in AI outcomes are critical. Fortune 500 companies with robust AI operations tend to incorporate elements of this technical report into their development processes.

Industry Leaders: Companies in finance, healthcare, and human resources have been at the forefront of adopting ISO/IEC TR 24027:2021. Tech giants like IBM, Microsoft, and Google have also integrated their principles into their AI development pipelines.

IEEE Ethically Aligned Design

The IEEE Ethically Aligned Design framework has gained traction, especially among companies focused on AI research and development.

Adoption Rate: While precise adoption rates are unavailable, the framework has been influential since its first edition release in March 2019. Its principles have been increasingly integrated into AI development practices across various industries.

Industry Leaders: Research-oriented AI companies and academic institutions have been particularly enthusiastic about incorporating the IEEE framework's principles into their AI development pipeline.

NIST AI Risk Management Framework (AI RMF 1.0)

The NIST AI RMF 1.0, released in January 2023, has seen rapid adoption, particularly among U.S.-based companies and those doing business with the federal government.

Adoption Rate: While specific adoption rates are not publicly available, the framework has gained significant traction since its release. Its integration into federal AI procurement processes, expected to begin in January 2024, will likely drive further adoption.

Industry Leaders: Companies involved in U.S. federal contracts, as well as those in highly regulated industries, have been early adopters of the NIST AI RMF 1.0.

Overall Insights on Adoption Trends

1. Regulatory Compliance: Organizations prioritize standards that align with emerging regulations, driving the adoption of comprehensive frameworks like ISO/IEC 42001:2023.

2. Sector-Specific Focus: Industries such as finance and healthcare are showing higher adoption rates for bias mitigation standards like ISO/IEC 23894:2023, given the critical nature of fair decision-making in these sectors.

3. Competitive Advantage: Early adopters of these standards use their compliance as a differentiator in the market, particularly when competing for government contracts or in highly regulated industries.

4. Integration Challenges: Many organizations are struggling to integrate these standards with rapidly evolving AI capabilities, leading to the development of more adaptive governance frameworks.

5. Cross-Standard Alignment: Leading companies are not adopting these standards in isolation but are creating integrated compliance programs that address multiple frameworks simultaneously.

6. SME Adoption Lag: While large tech companies are leading in adoption, small and medium-sized enterprises (SMEs) are lagging due to resource constraints and the complexity of implementation.

7. Geographical Variations: Adoption rates vary across regions, with countries like China showing increasing interest in international standards adoption to improve consistency with global practices.

Large tech companies and those in regulated industries leading the charge in adopting these standards. However, the dynamic nature of AI presents ongoing challenges in maintaining compliance while fostering innovation. As these standards continue to evolve, we can expect to see more adaptive and integrated approaches to AI governance across the industry. The following section provides more details on these frameworks.

A. International Standards

1. ISO/IEC 42001:2023 – AI Management System Standard

Key Dates:

- June 15, 2023: Standard published
- January 1, 2024: Certification process begins
- December 31, 2025: Expected widespread adoption deadline

Description:

ISO/IEC 42001:2023 establishes a comprehensive framework for AI governance, risk management, and lifecycle oversight. It provides organizations with a structured approach to manage AI systems, emphasizing continuous improvement and stakeholder trust.

Key Points:

1. *AI Governance Structure:*
   - Mandates clear roles and responsibilities for AI oversight
   - Requires board-level engagement in AI risk management

2. *Risk Assessment Methodology:*
   - Continuous risk identification and mitigation throughout AI lifecycle
   - Integration of AI risks into enterprise risk management frameworks

3. *Ethical AI Principles:*
   - Embedding fairness and non-discrimination in AI design and operation
   - Regular ethical impact assessments for AI systems

4. *Transparency and Explainability:*
   - Documentation requirements for AI decision-making processes
   - Mechanisms for providing meaningful explanations to stakeholders

5. *Continuous Monitoring and Improvement:*
   - Regular audits and performance evaluations of AI systems
   - Feedback loops for incorporating lessons learned into AI governance

Agentic AI-Specific Implications:

ISO/IEC 42001's emphasis on structured governance conflicts with Agentic AI's autonomous nature. While Agentic AI systems aim for self-governance and adaptation, the standard mandates:

- Human-centric oversight and predefined risk controls
- Detailed documentation of decision-making processes
  This creates significant challenges:
- Implementing governance structures that can keep pace with rapidly evolving AI behaviors
- Balancing autonomy with the need for human-understandable risk assessments

Organizations must develop adaptive governance frameworks that can dynamically adjust to Agentic AI's evolving capabilities while maintaining compliance with ISO/IEC 42001's structured approach.

2. NIST AI Risk Management Framework (AI RMF 1.0)

Key Dates:

- January 26, 2023: Framework released
- July 1, 2023: Implementation guidance published
- January 1, 2024: Expected integration into federal AI procurement

Description:

The NIST AI RMF 1.0 provides a comprehensive approach to identifying, assessing, and mitigating risks associated with AI systems throughout their lifecycle. It offers a flexible, non-prescriptive framework adaptable to various AI applications and organizational contexts.

Key Points:

1. *Governance Structure:*
   - Defining roles and responsibilities for AI risk management
   - Integration of AI risks into enterprise risk frameworks

2. *Risk Identification:*
   - Systematic approaches to AI-specific risk discovery
   - Stakeholder engagement in risk identification processes

3. *Risk Measurement and Assessment:*
   - Quantitative and qualitative risk assessment methodologies
   - Scenario analysis for emerging AI risks

4. *Risk Mitigation Strategies:*
   - Technical and procedural controls for AI risks
   - Continuous monitoring and adaptive risk management

5. *Transparency and Accountability:*
   - Documentation requirements for risk management decisions
   - Mechanisms for external audits and stakeholder communication

Agentic AI-Specific Implications:

NIST AI RMF's structured risk management approach conflicts with Agentic AI's dynamic risk landscape. While Agentic AI continuously evolves its capabilities and potential risks, the framework recommends:

- Predefined risk categories and assessment methodologies
- Static risk mitigation strategies and controls. This creates operational challenges:
  - Developing real-time risk assessment for self-modifying AI systems
  - Balancing innovation with consistent risk management practices

Organizations must implement adaptive risk management frameworks that can evolve alongside Agentic AI's capabilities while still meeting NIST AI RMF's comprehensive risk governance standards.

B. Industry-Specific Frameworks

1. Financial Services – Basel Committee AI Risk Management

Key Dates:

- January 2024: Initial guidelines published.
- December 2025: Full compliance expected for global banks.

Description:
The Basel Committee's framework addresses AI risks in banking, emphasizing robust risk modeling, fraud detection, and compliance automation. It applies to AI-driven systems in credit scoring, algorithmic trading, and anti-money laundering (AML).

Key Points:

1. *Risk Modeling:*
   - Agentic AI must validate risk models against historical financial crises.
   - Real-time stress testing for autonomous trading algorithms.

2. *Continuous Auditing:*
   - Automated audit trails for AI-driven transactions.
   - Real-time anomaly detection in high-frequency trading systems.

3. *Bias Mitigation:*
   - Fairness audits for AI-driven loan approvals.
   - Demographic parity checks in credit scoring models.

4. *Regulatory Alignment:*
   - Integration with EU AI Act and U.S. SEC rules.
   - Cross-border compliance for multinational AI deployments.

5. *Cybersecurity*:
   - Encryption of AI model weights in fraud detection systems.
   - Adversarial testing for AML algorithms.

Agentic AI-Specific Implications:
Basel's focus on static risk models conflicts with Agentic AI's adaptive decision-making. While Agentic AI systems optimize strategies in real-time (e.g., fraud detection), the framework mandates:

- Predefined validation benchmarks for risk models.
- Human review of AI-driven trading anomalies.
  This creates operational tensions:
- Delayed responses to emerging financial threats.
- Compliance costs for retrofitting adaptive AI to static audit requirements.
  Organizations must deploy hybrid systems where Agentic AI operates within Basel-approved risk boundaries while retaining limited autonomy for real-time adjustments.

2. Healthcare – FDA AI/ML Guidelines

Key Dates:

- October 2022: The FDA released guidance on distributed manufacturing and point-of-care manufacturing of drugs.
- March 2023: The FDA released guidance on artificial intelligence in drug manufacturing.
- April 2023: The FDA released guidance on submissions for AI/ML-enabled devices.
- May 2023: The FDA released an AI/ML for drug development discussion paper.
- March 2024: The FDA released guidance on considerations for using AI to support regulatory decision-making for drug and biological products.
- December 2024: The FDA released final guidance on predetermined change control plans (PCCPs) for AI/ML-enabled devices.
- January 2025: The FDA released a draft guidance on AI-enabled device software functions, which focuses on lifecycle management and marketing submission recommendations.
- January 2025: The FDA released draft guidance on the use of AI for decision making for drug and biological products Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products

Description:
The FDA's framework ensures safety and efficacy of AI/ML in healthcare, covering diagnostic tools, treatment recommendations, and patient monitoring systems.

Key Points:

1. *Clinical Validation:*
   - Agentic AI must demonstrate accuracy across diverse patient demographics.
   - Real-world performance monitoring for diagnostic algorithms.

2. *Post-Market Surveillance:*
   - Continuous reporting of AI-driven diagnostic errors.
   - Software updates tracked for algorithmic drift.

3. *Explainability*:
    - Clinician-interpretable rationale for treatment recommendations.
    - Audit trails for AI-driven patient triage decisions.

4. *Data Governance:*
    - HIPAA-compliant training data for AI models.
    - Model providence, where models originate from and how they were built
    - Patient consent protocols for AI-driven care plans.

5. *Interoperability*:
    - Integration with EHR systems without compromising performance.
    - Standardized APIs for multi-hospital AI deployments.

Agentic AI-Specific Implications:
FDA's requirement for static validation clashes with Agentic AI's self-improving capabilities. While Agentic AI in diagnostics evolves with new patient data, the guidelines demand:

- Fixed performance benchmarks pre-deployment.
- Human sign-off on material algorithm changes.
  This introduces implementation challenges:
- Delayed adoption of life-saving AI innovations.
- Resource-intensive revalidation for adaptive systems.

Healthcare providers must implement "version-locked" Agentic AI that pauses learning during FDA review cycles while maintaining baseline functionality.

3. Critical Infrastructure Protection – US Department of Homeland Security Safety and Security Guidelines for Critical Infrastructure Owners and Operators

Key Dates:

- November 2023: CISA and the United Kingdom's National Cyber Security Centre (NCSC) co-developed *Guidelines for Secure AI System Development*, setting security guardrails for AI system development.
- January 2024: The Cybersecurity and Infrastructure Security Agency (CISA) completed a *Cross-Sector AI Risk Analysis* based on sector-specific AI risk assessments conducted by Sector Risk Management Agencies (SRMAs).

- March 2024: The White House Office of Management and Budget (OMB) issued government-wide policy *M-24-10: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence*.
- March 2024: The U.S. Department of the Treasury published a report identifying AI-related cybersecurity and fraud risks in the financial services sector.
- April 2024: The Department of Homeland Security (DHS) published *Mitigating Artificial Intelligence (AI) Risk: Safety and Security Guidelines for Critical Infrastructure Owners and Operators* in response to Executive Order 14110.
- April 2024: The National Security Agency's (NSA) AI Security Center released a joint *Cybersecurity Information Sheet* on securely deploying AI systems, co-sealed by multiple international cybersecurity agencies.

Description:
The *DHS AI Safety and Security Guidelines for Critical Infrastructure* provide a risk-based framework for securing AI systems across 16 critical infrastructure sectors. Developed in response to Executive Order 14110, the guidelines integrate the NIST AI Risk Management Framework (AI RMF 1.0) and address AI-driven cybersecurity threats, operational risks, and supply chain vulnerabilities. As Agentic AI becomes more prevalent and introduces self-learning, autonomous decision-making systems, these guidelines establish baseline security measures while highlighting the challenges of governing AI that evolves post-deployment.

Key Points
1. *AI Risk Management for Critical Infrastructure*
    - Introduces a continuous risk assessment model for AI applications across 16 critical infrastructure sectors.
    - Establishes AI-specific sector risk assessments to address cybersecurity, operational reliability, and emergent AI threats.

2. *Cybersecurity Controls for AI Systems*
    - Requires proactive security monitoring of AI-driven systems to prevent adversarial manipulation.
    - Emphasizes secure AI model development practices, including robust authentication, anomaly detection, and adversarial defense techniques.

3. *Governance and Compliance Alignment*
    - Aligns AI security protocols with existing frameworks such as NIST AI RMF, ISO 42001, and sector-specific cybersecurity regulations.
    - Encourages cross-sector coordination between government agencies, private sector operators, and AI vendors to establish a unified AI risk management framework.

4. *AI Supply Chain Security and Procurement Standards*

- Establishes security guidelines for AI components, training data integrity, and AI supply chain risk management.
- Introduces minimum security requirements for AI vendors and third-party developers working with critical infrastructure sectors.

5. *Incident Response and AI System Resilience*
    - Mandates sector-wide reporting protocols for AI-related incidents, including adversarial attacks, system failures, and unanticipated AI behaviors.
    - Defines AI-specific recovery and continuity planning strategies, ensuring that AI-driven critical infrastructure systems remain operational during cyber incidents or adversarial compromise.

Agentic AI-Specific Implications:

The DHS AI Safety and Security Guidelines provide a structured approach to AI risk management, but Agentic AI's self-learning, autonomous nature introduces challenges that static compliance frameworks struggle to address. Traditional AI governance relies on predefined risk assessments, cybersecurity controls, and centralized oversight, but Agentic AI continuously evolves, adapts, and makes independent decisions, creating gaps in regulatory enforcement, security resilience, and incident response. To align Agentic AI with DHS's security guidelines, organizations must implement adaptive, real-time risk management strategies that account for AI autonomy, emergent behaviors, and decentralized decision-making.

- Fixed compliance benchmarks fail to capture Agentic AI's evolving decision logic, necessitating real-time AI risk monitoring solutions.
- Traditional security frameworks assume static attack surfaces, but Agentic AI introduces shifting vulnerabilities requiring self-healing defenses.
- Centralized governance models clash with multi-agent AI ecosystems, requiring federated AI governance structures.
- AI models that ingest external data post-deployment demand live integrity verification beyond static supply chain audits.
- Traditional forensic tools struggle to trace emergent AI behaviors, necessitating self-adaptive response mechanisms for real-time AI security incidents.

Additional Considerations

The DHS Safety and Security Guidelines for Critical Infrastructure Owners and Operators were developed in response to Executive Order 14110 and explicitly reference it multiple times as the directive mandating their creation. However, since Executive Order 14110 has been repealed, the legal and policy basis for these guidelines may need to be reconsidered.

Does the repeal of EO 14110 invalidate these guidelines?

- Not necessarily. While EO 14110 provided the initial directive for their creation, the guidelines incorporate frameworks from NIST AI RMF, CISA, and OMB AI policy M-24-10.

- DHS may still maintain and update them. The document explicitly states that DHS will "continue to update these guidelines" as AI risks and regulations evolve.
- They align with broader U.S. government AI security efforts. Many recommendations within the guidelines follow established cybersecurity best practices that are unlikely to be reversed just because EO 14110 is no longer in effect.

Key considerations:

- If a new Executive Order (EO 14141 or another) supersedes or contradicts these guidelines, DHS may need to revise them.
- Entities following these guidelines should monitor updates from DHS, NIST, and CISA for any modifications or new regulatory mandates.
- If your organization is leveraging these guidelines for compliance or risk management, consider mapping them against ongoing federal AI policy updates to ensure continued alignment.

4. HITRUST AI Security Assessment - Healthcare/Security

Key Dates:

- February 2024: HITRUST launches AI Security Assessment.
- March 2024: Initial adoption by early AI security adopters in healthcare and financial services.
- 2025: Expected broader industry adoption, including enterprise AI governance frameworks.

Description:

The HITRUST (Health Information Trust Alliance) AI Security Assessment is a new compliance framework designed to help organizations evaluate and mitigate AI-specific security risks. It provides up to 44 security controls tailored for AI platforms, focusing on risk management, compliance alignment, and governance. The framework allows organizations to leverage control inheritance, meaning companies can inherit compliance from cloud providers and third-party vendors instead of implementing security controls from scratch.

Key Points:

1. *AI Risk Management*
   - Establishes security requirements for AI-driven decision-making systems.
   - Emphasizes risk-based security controls for AI in sensitive environments like healthcare and finance

2. *Shared Responsibility & Control Inheritance*
   - Organizations can inherit compliance from cloud service providers, SaaS vendors, and AI model providers, reducing redundant security assessments.
   - Standardizes compliance efforts across AI ecosystems.

3. *Governance & Compliance Alignment*
    - Integrates with ISO 42001, NIST AI RMF, and existing healthcare compliance frameworks.
    - Supports regulatory mandates for AI security in HIPAA, GDPR, and financial services regulations.

4. *AI-Specific Cybersecurity Controls*
    - Covers AI model security, adversarial resilience, and runtime monitoring
    - Addresses supply chain risks in AI model training and deployment

Agentic AI-Specific Implications

The HITRUST AI Security Assessment introduces a structured compliance approach that may conflict with Agentic AI's dynamic, self-adaptive decision-making. While Agentic AI continuously learns and modifies its behavior, HITRUST's model requires:

- Predefined security controls that must be updated as AI models evolve.
- Explicit documentation and governance mechanisms for AI-driven actions, which may slow down real-time agentic decision-making.
- Formal compliance attestations that could limit AI autonomy in regulated sectors.

For more details, visit the official HITRUST website: HITRUST AI Security Assessment.

Comparative Analysis of AI Ethical Frameworks

Aspect	UNESCO AI Ethics Framework	OECD AI Principles	G7 AI Code of Conduct
Adoption Date	November 24, 2021	May 22, 2019	October 30, 2024
Number of Adopting Entities	193 member states	38 member countries	G7 nations
Key Focus Areas	Human Rights Protection Environmental Sustainability Equity and Inclusion	Transparency Accountability Inclusive Growth	Safety Reliability International Alignment

Aspect	UNESCO AI Ethics Framework	OECD AI Principles	G7 AI Code of Conduct
Human Rights Approach	Prohibits surveillance AI Mandates impact assessments for vulnerable groups	Emphasizes transparency Promotes digital literacy	Implements privacy protections Provides opt-out mechanisms
Technological Governance	Diverse stakeholder participation Public registries for high-risk AI	Comprehensive audit trails Legal liability frameworks	Red teaming requirements Kill switches for autonomous systems
Environmental Considerations	Carbon footprint reporting Energy-efficient algorithm promotion	Limited focus	Not a primary emphasis
Agentic AI Specific Challenges	Requires human oversight Mandates environmental impact disclosures	Demands static documentation Requires explainable AI	Predefined operational boundaries Human confirmation for novel approaches
International Collaboration	Technical assistance for developing nations Shared governance standards	Shared AI risk classification Cross-border incident protocols	Mutual AI certification recognition Shared safety research repositories
Accountability Mechanisms	Grievance redress systems Bias audits for public service AI	Lifecycle audit trails Workforce transition programs	Executive liability Third-party system auditing

Key Implications for Organizations

1. **UNESCO Framework**: Emphasizes holistic, human-centric AI development with strong social and environmental considerations.
2. **OECD Principles**: Focuses on transparency, accountability, and inclusive technological growth.
3. **G7 Code of Conduct**: Prioritizes safety, reliability, and controlled autonomous system development.

International Ethical Guidelines and Principles

1. UNESCO AI Ethics Framework

Key Dates:

- November 24, 2021: Framework adopted by 193 member states.
- January 1, 2023: Implementation guidelines released.

Description:
UNESCO's framework establishes global ethical principles for AI development, prioritizing human rights, equity, and environmental sustainability. It emphasizes inclusive governance and societal benefit, particularly for marginalized communities.

Key Points:

1. *Human Rights Protections:*
   - Prohibits AI systems enabling surveillance or social control.
   - Mandates impact assessments for AI's effects on vulnerable groups.

2. *Environmental Sustainability:*
   - Requires carbon footprint reporting for AI training.
   - Promotes energy-efficient algorithms for Agentic AI in climate-critical sectors.

3. *Equity and Inclusion:*
   - Diverse stakeholder participation in AI design.
   - Bias audits for systems used in education/public services.

4. *Accountability Mechanisms:*
   - Public registries for high-risk AI deployments.
   - Grievance redress systems for AI-harm victims.

5. *Global Cooperation:*
   - Technical assistance for developing nations.

- Shared standards for cross-border AI governance.

Agentic AI-Specific Implications:
UNESCO's human-centric principles clash with Agentic AI's autonomy. While Agentic AI aims for independent problem-solving, the framework mandates:

- Human oversight for systems affecting fundamental rights.
- Environmental impact disclosures for energy-intensive AI operations.
  This introduces operational friction:
- Delayed deployment due to multi-stakeholder governance requirements.
- Technical constraints on self-optimizing algorithms to meet sustainability targets.
  Organizations must implement ethical review boards to validate Agentic AI's alignment with UNESCO principles while maintaining operational efficiency.

2. OECD AI Principles

Key Dates:

- May 22, 2019: Principles adopted by 38 member countries.
- March 1, 2024: Updated guidelines for generative and agentic AI.

Description:
The OECD's principles promote trustworthy AI through transparency, accountability, and human-centric design. They serve as a foundation for national AI policies across member states.

Key Points:

1. *Transparency and Explainability:*
   - Public disclosure of AI system capabilities/limitations.
   - Real-time decision logs for autonomous systems.

2. *Robustness and Security:*
   - Adversarial testing for self-learning AI.
   - Fail-safe protocols for critical infrastructure AI.

3. *Accountability*:
   - Legal liability frameworks for AI-caused harm.
   - Audit trails covering entire AI lifecycle.

4. *Inclusive Growth:*
   - AI workforce transition programs.

- Digital literacy initiatives for underserved populations.

5. *International Collaboration:*
   - Shared AI risk classification systems.
   - Cross-border incident response protocols.

Agentic AI-Specific Implications:
OECD's transparency requirements conflict with Agentic AI's adaptive decision-making. While Agentic AI evolves dynamically, the principles demand:
- Static documentation of decision logic.
- Human-interpretable explanations for autonomous actions.
  This creates technical challenges:
- Developing explainability interfaces for neural networks that self-modify.
- Balancing performance optimizations with auditability needs.
  Organizations must deploy "explanation engines" that translate Agentic AI's complex operations into regulator-approved formats without compromising adaptability.

3. G7 AI Code of Conduct

Key Dates:
- October 30, 2024: Code announced at Hiroshima Summit.
- July 1, 2025: Voluntary adoption deadline for G7 nations.

Description:
The G7's code establishes ethical norms for advanced AI systems, focusing on safety, reliability, and international alignment. It targets generative and agentic AI in high-risk sectors.

Key Points:
1. *Safety Prioritization:*
   - Red teaming requirements for autonomous AI.
   - Kill switches for systems exceeding operational boundaries.

2. *Transparency Standards:*
   - Watermarking of AI-generated content.
   - Disclosure of training data sources/provenance.

3. *Privacy Protections:*
   - Differential privacy for self-improving AI.

- Opt-out mechanisms for AI-driven profiling.

4. *Global Alignment:*
   - Mutual recognition of AI certifications among G7 states.
   - Shared repositories for AI safety research.

5. *Accountability Measures:*
   - Executive liability for AI governance failures.
   - Third-party auditing requirements for critical systems.

Agentic AI-Specific Implications:

The code's safety-first approach challenges Agentic AI's exploratory nature. While Agentic AI thrives on unsupervised learning, the G7 mandates:

- Predefined operational boundaries for autonomous systems.
- Human confirmation for novel problem-solving approaches.
  This creates innovation bottlenecks:
- Restricted experimentation in dynamic environments (e.g., real-time crisis response).
- Increased compliance costs for multinational AI deployments.
  Organizations must implement "sandboxed autonomy" – allowing Agentic AI full independence within G7-approved risk corridors while maintaining override capabilities.

Acknowledgements

Contributors

Kayla Underkoffler, Zenity, State of Agentic AI Security and Governance Co-lead
Rock Lambros, RockCyber, LLC, State of Agentic AI Security and Governance Co-lead
Evgeniy Kokuykin, HiveTrace, State of Agentic AI Security and Governance Co-lead
Keren Katz, Tenable
Joshua Beck, SAS
Allie Howe, Cyber Growth
Ken Huang, DistributedSystems.AI, OWASP AVSS Co-lead
Sumit Ranjan, Forcespot
Vineeth Sai Narajala, Meta, OWASP AVSS Co-lead
Josh Devon
Victor Lu
Abhineeth Pasam
Kellen Carl
Ron Herardian
Ninad Doshi
Nayan Goel
Brian S Boyd
John Sotiropoulos ASI co-lead, Kainos
Ron F. Del Rosario ASI co-lead, SAP

Reviewers

Alejandro Saucedo - Chair of ML Security Project at Linux Foundation, UN AI Expert, AI Expert for Tech Policy, European Commission
Apostol Vassilev - Adversarial *AI Lead, NIST*
Chris Hughes - CEO, Aquia
Hyrum Anderson - CTO, Robust Intelligence
Steve Wilson - OWASP Top 10 for LLM Applications and Generative AI Project Chair and Chief Product Officer, Exabeam
Scott Clinton - OWASP Top 10 for LLM Applications and Generative AI Project Chair
Vasilios Mavroudis- Principal Research Scientist and Theme Lead, the Alan Turing Institute
Josh Collyer, Principal Security Researcher, Theme Lead
Egor Pushkin, Chief Architect, Data and AI at Oracle Cloud
Peter Bryan, Principal AI Security Research Lead- AI Red Team, Microsoft
Daniel Jones, AI Security Reseacrcher, AI Red Team, Microsoft
Michael Burgury, OWASP Low-Code/No-Code Lead, OWASP AIVSS Project Co-lead, Zenity

OWASP GenAI Security Project Sponsors

We appreciate our Project Sponsors, funding contributions to help support the objectives of the project and help to cover operational and outreach costs augmenting the resources provided by the OWASP.org foundation. The OWASP GenAI Security Project continues to maintain a vendor neutral and unbiased approach. Sponsors do not receive special governance considerations as part of their support.

Sponsors do receive recognition for their contributions in our materials and web properties. All materials the project generates are community developed, driven and released under open source and creative commons licenses. For more information on becoming a sponsor, visit the Sponsorship Section on our Website to learn more about helping to sustain the project through sponsorship.

Project Sponsors:

GenAI SECURITY PROJECT — GOLD PROJECT SPONSOR — TOP 10 for LLM and Generative AI

ByteDance · HIDDENLAYER · Mend.io · snyk · paloalto NETWORKS · TREND MICRO

GenAI SECURITY PROJECT — SILVER PROJECT SPONSOR — TOP 10 for LLM and Generative AI

splx AI · Calypso AI · Acuvity · pangea · Cobalt

LAKERA · securiti · ActiveFence · Prompt: · TROJ.AI

Lasso SECURITY · PROTECTO · PROMPTARMOR · Synack · GT cyber

Sponsor list, as of publication date. Find the full sponsor list here.

Project Supporters

Project supporters lend their resources and expertise to support the goals of the project.

Accenture
AddValueMachine Inc
Aeye Security Lab Inc.
AI informatics GmbH
AI Village
aigos
Aon
Aqua Security
Astra Security
AVID
AWARE7 GmbH
AWS
BBVA
Bearer
BeDisruptive
Bit79
Blue Yonder
BroadBand Security, Inc.
BuddoBot
Bugcrowd
Cadea
Check Point
Cisco
Cloud Security Podcast
Cloudflare
Cloudsec.ai
Coalfire
Cobalt
Cohere
Comcast
Complex Technologies
Credal.ai
Databook
DistributedApps.ai
DreadNode
DSI
EPAM
Exabeam
EY Italy
F5
FedEx
Forescout
GE HealthCare
Giskard
GitHub
Google
GuidePoint Security
HackerOne
HADESS
IBM
iFood
IriusRisk
IronCore Labs
IT University Copenhagen
Kainos
KLAVAN
Klavan Security Group
KPMG Germany FS
Kudelski Security
Lakera
Lasso Security
Layerup
Legato
Linkfire
LLM Guard
LOGIC PLUS
MaibornWolff
Mend.io
Microsoft
Modus Create
Nexus
Nightfall AI
Nordic Venture Family
Normalyze
NuBinary
Palo Alto Networks
Palosade
Praetorian
Preamble
Precize
Prompt Security
PromptArmor
Pynt
Quiq
Red Hat
RHITE
SAFE Security
Salesforce
SAP
Securiti
See-Docs & Thenavigo
ServiceTitan
SHI
Smiling Prophet
Snyk
Sourcetoad
Sprinklr
stackArmor
Tietoevry
Trellix
Trustwave SpiderLabs
U Washington
University of Illinois
VE3
WhyLabs
Yahoo
Zenity

Sponsor list, as of publication date. Find the full sponsor list here.