

Securing the Agentic Era: The Action Blueprint for Agentic Security

The policy urgency

As the capabilities of frontier AI models rapidly progress and AI agents gain unprecedented reasoning, planning, and tool-use capabilities, it is urgent that we explore proactive and coordinated policy approaches for agentic safety and security, especially when security is also a fundamental driver of economic competitiveness. Ensuring we have a secure-by-design agentic ecosystem will foster the trust required for the widespread adoption and scaling of AI agents. This will in turn enable us to transform the [defender's advantage](#) – reversing the asymmetric burden of security – into an economic advantage. Building upon our [TRUST Framework](#) for broad agent governance, this paper aims to go one level deeper to outline concrete steps we can take to secure the agentic ecosystem. The Action Blueprint for Agentic Security is an action-oriented, technology-neutral blueprint designed for various stakeholders in the ecosystem to align governance, public procurement, technical standards, and defensive engineering.



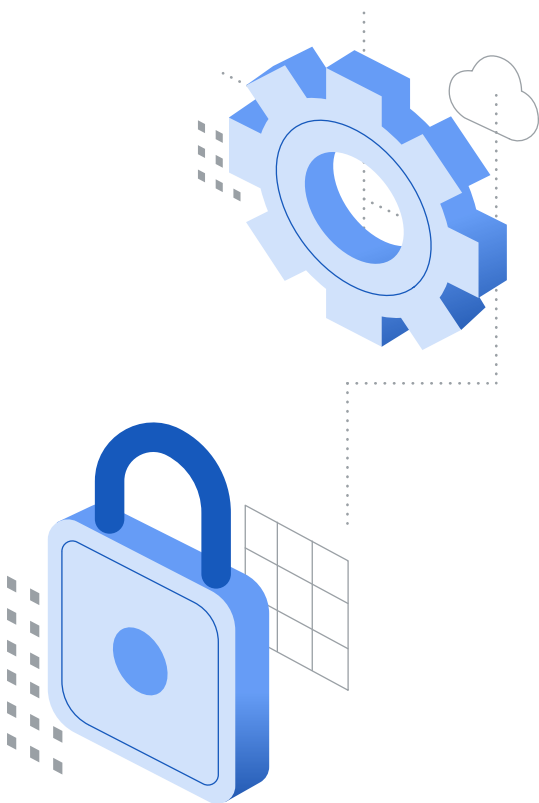
While general-purpose and specialized agents offer immense promise, they introduce [novel](#) threat vectors that challenge traditional security paradigms. Threat actors can increasingly [automate](#) the discovery and exploitation of software vulnerabilities – including those in widely used open-source libraries – at a speed and scale that outpaces manual security patching. Simultaneously, the rise of “[vibe-coding](#),” where users deploy custom agents to build and run high-privilege tools, often without awareness of cybersecurity best practices, creates significant operational risks inside enterprise and personal networks.

As a result, new [threat vectors](#) targeting agents and multi-agent systems are emerging, presenting novel risks to the agentic ecosystem at a time when we are still building an understanding of multi-agent behavior and related cyber risks. A [DarkReading poll](#) found that nearly half of cybersecurity professionals believe agentic AI will represent the top attack vector for cybercriminals and nation-state threats in 2026. In parallel, as model capabilities continue to advance, we are starting to observe AI agents exhibiting unexpected and sometimes coordinated adversarial behavior in order to achieve testing objectives during evaluation. These unexpected behaviors already have real-world impact and consequences, despite the lack of malicious intent.

These developments highlight the mounting need to both rapidly secure AI development and deployment infrastructure and widely deploy defensive agents. It has become critical to automate cyber defense as much as possible, because manual security operations are woefully insufficient to counter the speed and scale of attacks by malicious threat actors using agents or by agents themselves.

Deploying agents for security offers the ecosystem a historic opportunity to [flip the traditional “defender’s dilemma”](#), which states that defenders have to get things right 100% of the time, while the attacker (whether human or agent) only has to find one gap. For example, by shifting security operations to “machine speed,” we can use defensive AI agents to autonomously seek out and [eradicate vulnerabilities](#) before they can be exploited or even committed to code. Google’s [CodeMender](#) defensive agent operates like a senior security researcher that never sleeps, identifying vulnerabilities and writing then applying high-quality, verified code fixes at scale. By pairing autonomous discovery with automated remediation, we are transitioning from a reactive posture of patching bugs to a proactive posture of engineering secure-by-design systems. However, we can only capitalize on this advantage through deployment and adoption of agentic defense tools.

Securing the agentic era requires public and private sector leaders to establish proactive security approaches before a crisis occurs. By building on existing security frameworks, such as the NIST Cybersecurity Framework (CSF) and Google’s Secure AI Framework (SAIF 2.0), the Action Blueprint for Agentic Security introduces agent-specific guardrails, including secure-by-design agentic protocols and autonomous defensive patching, to ensure agentic AI technology develops with security as a priority.



The Action Blueprint for Agentic Security

We believe that a safe, reliable, and secure agentic ecosystem is the foundation of sustainable, widespread adoption, which will enable society to fully reap the benefits of AI agents. The following section details 11 core recommendations of the Action Blueprint, calling on governments, standards bodies, industry leaders, and developers to take decisive action to secure the agentic ecosystem.

Standards and protocols

Design and adopt secure-by-design agentic protocols and create agent security standards.

Standards Developing Organizations (SDOs) should prioritize the standardization of definitions for agentic concepts and terms and core protocols (e.g. [A2A/AP2/UCP](#)) to create a foundational layer for interoperability and security. As more organizations leverage agents internally, security leaders should transition their enterprise security architecture to approaches such as [Beyond Zero](#), a contextual, risk-based, resource-level authorization model that can run at machine speed.

Evaluation and governance

Raise the security baseline for model testing and evaluations.

Organizations should treat the testing and evaluation of AI models as a type of deployment. This can include ensuring secure configuration of testing infrastructure, enhanced security for evaluation environments, increased monitoring during testing and of evaluation results, and a defense-in-depth approach to model testing.

Establish pilot programs and sandboxes for security use cases.

Governments should fund pilot programs for public sector cybersecurity use cases to demonstrate and evaluate the safe and beneficial use of agents. For example, governments and sector risk management agencies can create secure, controlled [sandboxes](#) to allow for research, development, and testing of agentic cybersecurity systems, including for critical infrastructure environments, to accelerate innovation while gathering evidence to assess future policy decisions.

Champion adaptable, technology-neutral governance.

Governments should review and evaluate existing legal frameworks, including sectoral regulations, to ensure they remain outcome-focused and technology-neutral in the age of agents. This approach avoids imposing constraints on the underlying, evolving technology and provides flexibility for iteration and improvement as real-world insights are gathered.

Collaboration and information sharing

Share information and collaborate on agentic AI security and safety practices.

We invite cross-industry and cross-sector partners to engage through existing global forums such as the [Coalition for Secure AI](#), [Frontier Model Forum](#), and [ML Commons](#) to ensure that security breakthroughs can iteratively improve ecosystem resilience. Further, we should work together to ensure incident reporting standards include as much relevant information as appropriate to empower the nascent field and science of agent security incident forensics.

Upgrading infrastructure and code

Modernize digital public infrastructure for AI agents.

Governments should prepare public sector digital services, including citizen-facing portals and core infrastructure, to securely receive and interact with increasing numbers of AI agents. This includes adopting cryptographically verifiable [Agent Identities](#) and open, standardized interoperability protocols like [A2A \(Agent-to-Agent\)](#) to reliably authenticate trustworthy agents, scope their data access, and distinguish them from malicious bots. This could also include establishing operational thresholds and tools to mitigate the risk of runaway agent coordination.

Integrate AI agents into secure code development and transition processes.

Currently, many critical software vulnerabilities stem from memory safety exploits in legacy code. AI agents can increasingly support translating this code into memory-safe languages (like Rust). Governments should explore funding Focused Research Organizations to systematically identify, translate, and verify the most heavily relied-upon and under-resourced open-source libraries into memory-safe languages using agents. Separately, AI agents can already help us generate secure code by default, catching bugs before they even enter the codebase. Organizations should consider integrating AI agents into existing software development processes to enhance secure code development.

Resilience and readiness

Prepare for the rapid escalation of discovered vulnerabilities and novel agentic threats.

Security leaders in critical sectors should prepare for threats such as prompt injection, rogue actions, unintended agent coordination, agent-to-agent social engineering, and identity spoofing by identifying short-term incident response and vulnerability patching plans, implementing industry best practices in the medium-term, and establishing long-term resilience.

Protect ourselves and the wider agentic ecosystem by adopting baseline security hygiene.

Developers and users should adopt controls laid out in the [Secure AI Framework 2.0](#), especially when deploying “persistent-by-default” high-privilege agent configurations. Resources such as Google’s [Agent Security Assessment](#) can help.

Enforcing Secure-by-Design

Promote Secure-by-Design systems through procurement requirements.

Use public procurement as a mechanism to drive market demand for secure and trustworthy agents. By embedding security and privacy-by-design requirements into public sector contracts, governments can raise the security baseline for the entire agent ecosystem. For example, governments can request vendors to attest to established frameworks like the NIST Secure Software Development Framework (SP 800-218) and its AI profile (SP 800-218A). Furthermore, procurement contracts can stipulate that vendors demonstrate compliance with foundational management systems like ISO/IEC 42001, while future-proofing deployments by aligning with technical standards such as ISO/IEC 27090 (FDIS) for AI lifecycle cybersecurity.

Utilize defensive agents to conduct automated vulnerability discovery and patching.

Coordinate across the ecosystem to deploy agentic cybersecurity tools, such as [CodeMender](#), for automated vulnerability discovery and patching, especially by critical infrastructure providers. This includes providing funding, training, and access to tools to ensure system administrators at critical facilities (e.g., hospitals, power grids, water utilities) have access to AI agents for cyber defense.

The rise of AI agents is a major opportunity to redefine how we interact with technology, but success depends on building and maintaining trust. At Google, we bring our decades of experience securing the internet at scale to this new challenge. Whether you are a developer creating new applications, a security leader defending critical infrastructure, a peer developing frontier models, or a policymaker shaping the government's response, we invite you to join us. By deploying agents to autonomously discover and patch vulnerabilities, committing to foundational security hygiene, increasing multi-layered security for model testing and evaluation, and establishing clear, interoperable standards, we can build an agentic future that is not only profoundly capable, but secure by design. Let's build a safer, resilient agent ecosystem, together.

