June 2026

The Three Layers of Agent Security

A Framework for Policymakers

Shaun Ee and Pegah Maham

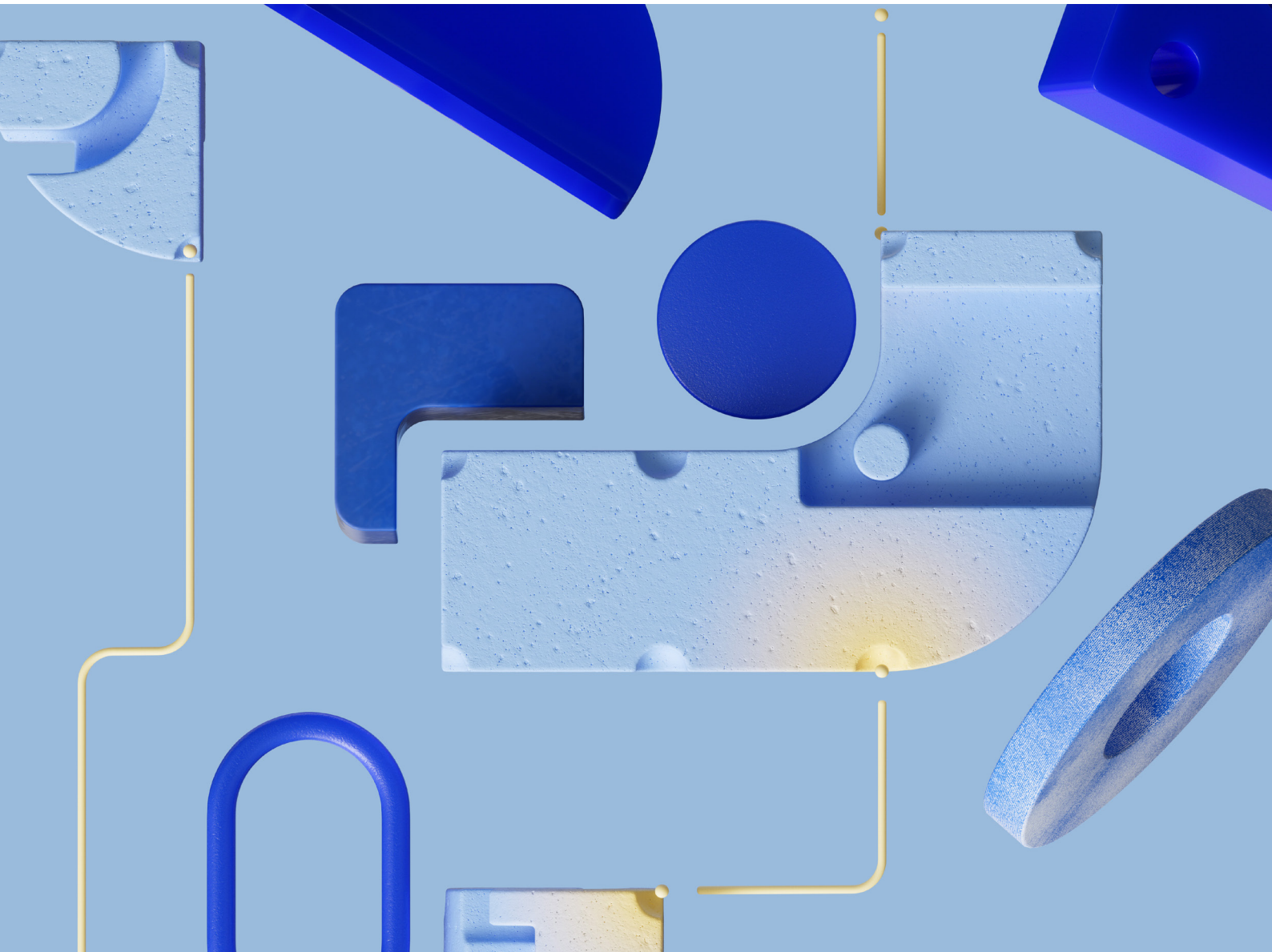

Table of Contents

Executive summary

Executive summary

AI technology is moving from AI chatbots toward AI agents—systems that combine the intelligence of advanced AI models with access to digital tools, empowering the agents to take actions on behalf of users, under their control. Beyond enabling personalised digital assistants, the advent of AI agents represents a larger, structural shift toward AI systems becoming active participants in the digital economy. If realised, this transition has massive economic potential, with McKinsey estimating that redesigning workflows around human-agent partnerships could unlock $2.9 trillion in economic value in the U.S. alone by 2030. At the same time, this structural shift introduces a complex web of new challenges for policymakers.

Digital security, the focus of this white paper, is among the most immediate hurdles. Because AI agents can act autonomously across interconnected platforms, a lack of proper safeguards could create sprawling new attack surfaces while simultaneously giving malicious actors powerful new tools. To realise the potential of AI agents, we need to defend these novel attack surfaces, govern dual-use risks, and ensure defenders maintain an edge by using frontier AI models and new tools. Addressing these security risks from a structural perspective requires a three-layered approach:

Part I: Securing individual AI agents. Securing the AI agent ecosystem requires a comprehensive, multi-layered approach that addresses the unique risks posed by agent autonomy, such as sensitive data disclosure, rogue actions, and manipulation. Security must be integrated throughout the development lifecycle, from model-level interventions like rigorous training and hardening to system-level architectures that include input filtering, behavioural monitoring, and oversight. Policymakers and industry should work together to drive the adoption of standards and best practices on agent security, including around agent definitions and core functional protocols, so as to provide a bedrock of interoperability and security for the ecosystem.

Two avenues of pre-standardisation work can help lay broader foundations for secure and widespread AI adoption:

- **Consensus-building:** Better clarity on desired objectives can make trade-offs clearer, and improve technical design choices. It can also improve coordination across different standards developers, avoiding a fragmented landscape of overlapping frameworks. AI Security Institutes (AISIs), like the U.S. Center for AI Standards and Innovation (CAISI) and the UK AI Security Institute (AISI), can facilitate this consultative process, such as through CAISI's January 2026 Request for Information on AI Agent Security.

- **R&D:** Without adequate exploration of the design space for agent security, we risk prematurely locking in suboptimal choices for standardisation. Policymakers can facilitate R&D by creating space for industry and academics to innovate, such as with regulatory sandboxes, and via funding for academic research.

To ground these pre-standardisation efforts in practical applications, it helps to examine the integration of AI agents through the lens of traditional enterprise security. By exploring three illustrative domains, we can more concretely identify how cybersecurity principles apply to AI agents, and where consensus-building and R&D can adapt them for the agentic era.

- **Identity and access management (IAM):** As practitioners adapt IAM solutions for the agentic era, cross-industry consultation and field trials (e.g., in finance and healthcare) could clarify how to balance the security benefits of agent identifiers against privacy, usability, and sector-specific needs. R&D on privacy-preserving technologies could also make it easier to reap these security benefits while protecting sensitive data.

- **Secure software development practices:** To help downstream developers secure their agentic applications, industry best practices and guidelines that emphasise secure-by-default and secure-by-design approaches can help create a consistent baseline for security against accidents or efforts by bad actors trying to subvert or redirect agents' activities. Further R&D by industry could also help to develop state-of-the-art defences and refine them for use by downstream developers in production environments.

- **Monitoring and detection:** To develop ways to overcome the limitations of human oversight at scale, automated "AI monitors" to oversee AI agents are emerging. Further R&D and field trials could help identify how accurate, cost-efficient, and reliable such AI monitors are.

Part II: Addressing multi-agent security risks. Multi-agent dynamics can enable de-centralised, complex task automation and coordination, such as scientific research. At the same time, as AI agents increasingly collaborate and interact within complex "virtual agent economies," they can introduce risks—such as cascading failures, tacit collusion, and accountability vacuums—that individual agent safeguards cannot necessarily fully mitigate. These environments are also susceptible to large-scale, synchronised attacks like "systemic traps," which might exploit the homogeneity and interconnectedness of networks of agents to cause widespread disruption. Addressing these challenges requires a shift toward an interdisciplinary, systems-governance approach that draws on, for example, economics and cybersecurity. By prioritising research into intelligent delegation protocols, trust and reputation architectures, and empirical threat modelling, stakeholders can establish the market infrastructure and "rules of participation" necessary to ensure that collective agent behaviours remain accountable and secure.

To address these issues, multi-agent safety can benefit from a layered, systems governance approach. **Interdisciplinary research straddling AI, cybersecurity, economics, and other fields could boost our confidence in how to govern a world with many interacting AI agents:**

- **Understanding multi-agent risks:** To build understanding and consensus around which multi-agent risks might deserve the most attention, AISIs and researchers could conduct interdisciplinary, empirically informed research that taps not just on existing AI literature but also cybersecurity, economics, and other disciplines.

- **Piloting market infrastructure:** As these virtual agent economies develop, researchers could explore architectural elements to safely manage how AI agents interface with one another and the real-world economy, such as intelligent delegation protocols to ensure accountable relationships between AI agents.

- **Experimenting with market design:** In the longer term, researchers could investigate even more experimental market design schemes, such as a trust and reputation architecture to scalably govern

multi-agent ecosystems. To catalyse cross-disciplinary research, government science foundations and private philanthropies could sponsor grand challenges to address problems in this space, and perhaps consider regulatory sandboxes to stress-test such measures before expanding them to the wider economy.

Part III: Empowering cyber defenders and building resilience. The rise of AI agents introduces a dual-use challenge in cybersecurity, where attackers can leverage agentic capabilities to increase the speed and scale of threats, but defenders can similarly transform their security posture through automated vulnerability remediation, adaptive threat detection, and enhanced incident response.

To ensure that society can meet current cybersecurity risks presented by frontier AI capabilities, defenders will need to invest in a wide range of efforts. Policymakers, industry, and academia should collaborate to advance initiatives across five main goals:

First, strengthen understanding of the threat landscape by monitoring progress in AI capabilities and adoption by threat actors. AISIs, frontier AI developers, and academics can invest in building novel advanced cyber capability evaluations, particularly realistic multi-stage attack scenarios; intelligence agencies can also monitor adoption by threat actors, such as by treating the adversarial use of AI as an intelligence collection priority.

Second, software developers can secure their products and services by improving software product security and supply chain integrity:

- **Improve the security of software products:** Software vendors should adopt a secure-by-design approach as laid out in guidance from national cyber agencies, and procurement agencies can incentivise adoption via standards. Meanwhile, industry and academia can conduct research on methods for making AI-generated code more secure, while standardised benchmarks for code security can encourage a "race to the top."

- **Fortify supply chain integrity:** Software developers should adopt frameworks like Google's Supply-chain Levels for Software Artifacts (SLSA) and Wiz's Software development lifecycle Infrastructure Threat Framework (SITF) to secure the software supply chain. To support the open-source ecosystem, industry, research agencies, and cyber agencies can make adoption easier via tooling, funding, and compute.

Third, organisations can defend digital systems in deployment by proactively hardening their deployed systems and adopting AI agents for security operations:

- **Reduce the attack surface:** Where possible, organisations should use AI as a force multiplier to identify weaknesses and fix system security issues (e.g., configuration, automated asset discovery/remediation). However, some systems cannot be fixed easily. Legacy systems should be deprecated and replaced. Where critical infrastructure cannot be patched or replaced, public or private R&D funders can use pull mechanisms to incentivise new AI-enabled tools that enable automated hardening and compensating controls, such as anomaly detection or airgapping.

- **Adopt AI agents for security operations:** Organisations should use agentic AI to automate

workflows like triage and containment. To scale these defences, governments can fund adoption, fast-track procurement, and create regulatory sandboxes for critical sectors. Developers, deployers and researchers must also build safeguards to prevent these defensive tools from being exploited by adversaries.

Fourth, empower and mobilise defenders by investing in tools, access, and talent to achieve the above goals:

- **Invest in defensive AI tools:** To provide a diverse stack of AI-cyber capabilities usable at a wide range of price points, frontier AI developers can build a broad spectrum of models, which they and third parties can build further scaffolding on top of, as Google DeepMind has done with CodeMender. To spur innovation in areas that the commercial market might neglect, research funders and cyber agencies can identify gaps through consultation, and use pull mechanisms to incentivise new tool development.

- **Bolster talent and institutional capacity:** Governments can provide surge funding, training, and capacity for under-resourced public sector entities and critical infrastructure, and rally them to act by convening a 90-day government-wide sprint on national resilience. Industry, policymakers, and philanthropies can also design public service programmes to hire and train national teams of cyber experts, such as Google's work with the Consortium of Cybersecurity Clinics, or—at a larger scale—a "Security for America" programme in the spirit of Tech Congress and AmeriCorps.

Fifth, improve collective visibility into real-time threats by bolstering detection and information sharing. Frontier AI companies, cloud providers, and other infrastructure providers (e.g., telecoms, payments) can jointly explore shared security signals that enable detection of malicious agentic campaigns while preserving user privacy. While more consultation is needed, possible venues for such sharing could include the FMF or an independent "Agentic Security Exchange," drawing inspiration from the Global Signal Exchange that Google set up in 2024.

Background on AI agents

Background on AI agents

Generative AI systems respond only when prompted, primarily as a collaborative sounding board. AI agents, by contrast, can increasingly act as proactive assistants, delegates, and proxies. This capacity stems from a set of complementary capabilities, the most important being foundational improvements in the reasoning engines of AI.

AI technology is moving from AI chatbots toward AI agents—systems that combine the intelligence of advanced AI models with access to digital tools, empowering the agents to take actions on behalf of users, under their control. Beyond enabling personalised digital assistants, the advent of AI agents represents a larger, structural shift toward AI systems becoming active participants in the digital economy. If realised, this transition has massive economic potential, with McKinsey estimating that redesigning workflows around human-agent partnerships could unlock $2.9 trillion in economic value in the U.S. alone by 2030. At the same time, this structural shift introduces a complex web of new challenges for policymakers.

Agents can now plan over longer time horizons, an ability that has its roots in the paradigm shift from "instant reply" systems to AI systems that take time to think using chain-of-thought reasoning and inference-time scaling. Through reasoning and self-reflection, agents can take high-level goals, break them down, and iterate through multi-step workflows until an objective is met. The cognitive engine of AI agents leverages features and supporting infrastructure that include:

- Integration with external software, or "tool use," to let agents reliably manipulate their environments (such as querying databases, executing code or navigating web browsers);
- Expanded memory and context windows to keep focus during complex tasks;
- Multimodal perception and output to understand more than text; and
- Multi-agent orchestration frameworks to enable coordination with—and delegation to—other specialised agents; and protocols to standardise interfaces such as the Model context protocol (MCP) that standardises how applications expose tools and context to language models.

Combined, these capabilities have enabled a shift from answering to acting, which is unlocking breakthroughs across multiple domains. Google has built a Co-Scientist, which acts as a virtual scientific collaborator to formulate novel hypotheses and simulate experiments. Already, Co-Scientist has helped scientists to fight liver fibrosis, fast-track genetic leads to reverse cellular ageing, and collaborate on new approaches to ALS. Google's AlphaEvolve, a Gemini-powered coding agent to design advanced algorithms, has helped to suggest quantum circuits with 10x lower error than previous baselines, and to help resolve open mathematical challenges, including some of the elusive, decades-old Erdős problems.

Nor is scientific research the only area that AI agents are transforming. In healthcare, systems like Google's AMIE and Basalt Health's AI-powered assistants are preparing patient charts, driving clinical interviews, and analysing medical imagery to support complex diagnoses. And in commerce, enterprise agents are used to optimise customer interactions.

As agentic capabilities mature, the digital landscape will evolve from isolated applications into a rich,

interconnected ecosystem. For the most part, we are currently in the era of single agents, with individual systems, like solitary diagnostic assistants, deployed to execute specific, bounded tasks. Such systems sometimes comprise multiple agent instances integrated into a larger team: the Co-Scientist, for example, combines a range of specialised agents into a committee where agents build off each other's work (see Figure 1).

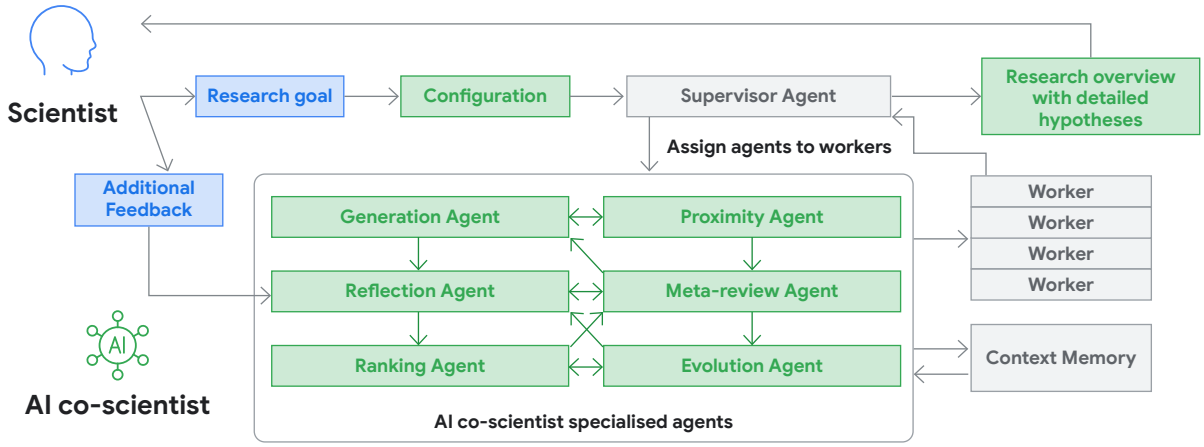

Figure 1: *Co-Scientist system overview. specialised agents (green boxes, with unique roles and logic); human scientist input and feedback (blue boxes); system information flow (dark gray arrows); agent-to-agent feedback green arrows within the agent section).*

As the web of agent interactions grows more complex and interlinked, it might in time lead us to "virtual agent economies,"as we discuss later in this paper: decentralised infrastructures where agents autonomously interact, negotiate, and transact with other agents across organisations. Beyond pure productivity, these multi-agent ecosystems offer novel ways to approach complex coordination bottlenecks in fields ranging from logistics to cyber defence. At their best, high-fidelity AI agent proxies could even amplify human agency by enabling "Coasean bargaining at scale," where agents acting as our faithful representatives in negotiating the frictions of everyday life could reduce negotiation costs to near-zero, turning previously unsolvable zero-sum dynamics into newly feasible, mutually beneficial arrangements.

In the future, companies may also build multi-agent bureaucracies that resemble entire organisations rather than individual teams. Since highly capable AI agents could communicate and learn rapidly, for example by copying, merging, and distilling their knowledge, some economists have suggested that agents could bypass human coordination bottlenecks that have historically limited firm size. However, they would also face significant challenges: for example, correlated errors across similar and widely used AI agents could make their parent organisations uniquely fragile to disruption.

Realising the full potential of AI agents, from individual use cases to the economy at large, will require looking beyond pure capability benchmarks. Just as is true of digital systems generally, security is an essential enabler for adoption. For users to trust AI agents to act on their behalf, they will need to feel confident that their data is protected and their agents will not take unexpected actions. As we explore below, building this resilient ecosystem must start with the most foundational element: individual agents themselves.

Part I: Securing individual agents

Part I: Securing individual agents

Securing the agent ecosystem begins with securing individual agents.[1] Agents have value because of their autonomy, flexibility, and capacity to interact with external tools and data to achieve longer-term goals: for example, a clinician could dispatch an AI agent to retrieve, analyse, and amend medical records without having to specify where these fragmented records are stored, or even an initial clinical suspicion.

Agents as a new attack surface

However, this also introduces new security risks relative to traditional software, as indicated in Google's Approach for Secure AI Agents. One major threat is sensitive data disclosure: agents have access to large volumes of organisational data, so malicious actors can trick them into leaking sensitive information like confidential patient files. Another is the agent executing rogue actions: that is, actions that are unintended by the user, harmful, or violate policies. Rogue actions exploit the trust placed in AI agents: for example, without appropriate security measures in place, a compromised AI agent could be used to authorise fraudulent medical billing without human oversight, or accidentally delete critical allergy information.

Both rogue actions or sensitive data disclosures can occur without malicious intent, stemming instead from misinterpretations or other accidental actions by the agent. The agent might misunderstand ambiguous instructions or context. Such cases involve harmful divergence from user intent due to the agent's interpretation, not external compromise. Additionally, unexpected negative outcomes can arise if the agent misinterprets complex interactions with external tools or in new environments. For example, it might misinterpret the function of buttons or forms on a complex website, leading to accidental purchases or unintended data submissions when trying to execute a planned action.

In the case of intentionally malicious manipulations, as AI agents expand in use and capabilities, attackers may deploy increasingly subtle and insidious attacks, as Google DeepMind's work on AI Agent Traps discusses. Via epistemic corruption, attackers could poison an agent's long-term memory or retrieval knowledge bases, leading it to give confidently wrong answers without ever executing a visibly rogue action. Attackers could also target the human-AI handoff, such as by having a compromised agent generate large volumes of plausible-looking evidence that causes human reviewers to rubber-stamp dangerous decisions.

As outlined in Google's Approach for Secure AI Agents, neither traditional software controls nor AI-based controls are fully sufficient to counter these threats. Deterministic security controls can be too restrictive for agents, but relying solely on a single model's reasoning can leave it vulnerable to mistakes or attacks. Balancing the risks and benefits of autonomy requires a new field of agent security. This is a team effort that has to involve all parts of the ecosystem, including both model providers and downstream developers and users.

Upstream agent security interventions

As a first line of defence, interventions by AI agent developers can help reduce the vulnerability of

[1] Malicious use of agentic cyber capabilities is discussed in Part III.

agents to potential attack vectors, and guard against undesired actions, as outlined in the Secure AI Framework. The reasoning engines of individual agents are vulnerable to LLM-based attacks, such as prompt injection and data poisoning. Agents may also make mistakes or be misaligned with the user's intent, such that they take undesirable actions even without external interference.

Google applies security interventions at both the model level, while training the LLMs that form the reasoning core of agents, and at the system level, when building the surrounding architecture to manage how the agent interacts with users and tools. Many of these model-level interventions are implemented during the "post-training" phase of model development. Via techniques like supervised fine-tuning and reinforcement learning from human feedback, models learn to follow instructions and refuse unsafe requests. Model hardening helps build resilience against threats like indirect prompt injection, using automated red teaming to generate realistic attacks that the model learns to recognise and refuse.

At the system level ("outside the model" interventions), Google implements a comprehensive suite of safeguards. This begins with protections like input and output classifiers that filter out harmful attacks, extending into a comprehensive defence-in-depth strategy that applies traditional cybersecurity concepts to AI, as described in Google's blog post on mitigating prompt injection attacks.

Downstream interventions to secure agentic applications

Part of the richness of AI agents stems from their adaptability. LLMs can be fine-tuned on additional data by downstream developers and integrated into other custom systems and applications, helping make them generally useful across domains. Fine-tuning can also affect safety features of the upstream model, even when the additional data has no harmful content. Integration with other tools and systems can then make an agent vulnerable to new attacks: "memory poisoning" attacks, for example, can implant malicious goals in an agent's long-term memory that persist over extended periods. Upstream developers do not have direct influence over these implementation aspects, so downstream developers play an important role in designing their own additions to agentic systems safely, and evaluating whether their changes might affect existing safety features.

For example, to prevent rogue actions, downstream developers should ensure that the tools an agent can access follow the principle of least privilege. If a developer wants its AI agent only to help clinicians find and summarise patient data, it should not have the authority to alter that data. And if the stakes are higher, as where clinicians do want AI agents to alter patient data, then decision-making should be structured so that users have control and oversight proportionate to risk. Increasing autonomy may make it harder for humans to keep track of all automated decisions, meaning that downstream developers should ensure that they have strong checks in place to monitor for and interrupt rogue agent actions where needed.

To guide downstream developers across these scenarios, Google has laid out three principles: agents must have well-defined human controllers, agent powers must have limitations, and agent actions and planning must be observable. These complement best practices laid out by other companies, like Meta's "Agents Rule of Two," which states that agents should at most have only two of three properties: the ability to process untrustworthy inputs; access sensitive systems/data; or execute binding actions or

communicate externally. These principles are just a start, and cooperation and research will be needed across the ecosystem to further advance best practices for agent security.

As outlined in Google DeepMind's AI Control Roadmap, for anticipated future capabilities, safety and security teams have started developing a holistic approach to defend against the impact of internal AI agents acting in unintended ways, in addition to reducing such behaviour in the first place. AI control is an approach to mitigating risk from agentic deployments focusing on system-level mitigations such as automated monitoring and response, access controls and incident response. Mirroring the security mindset from cybersecurity, the Control Roadmap conservatively treats certain internally deployed AI agents as untrusted, aiming to build defences that do not depend on trust that an internal agent will behave as intended.

A core component and opportunity of AI control is the use of chain-of-thought monitoring, an area where Google DeepMind has previously collaborated with the UK AI Security Institute (AISI) and others. By evaluating an agent's intermediate reasoning steps before it executes an action, it is possible to detect rogue actions more effectively and prevent incidents.

As AI agents take increasingly long sequences of autonomous actions, safeguards will need to look beyond individual actions, which can each appear harmless in isolation, to evaluate the potential harm of an action sequence as a whole and, eventually, correlate potential issues across agent sessions. Google DeepMind's Control Roadmap proposes a path for Google and other companies to not only better understand these issues, but also to mitigate them at scale.

Security standards as a foundation for agent adoption

Industry best practices and standards can lay the foundation for security in the agentic era. Without standards that are widely accepted and adopted, developers and deployers of AI agents must build security frameworks from scratch, which slows down progress, creates compatibility and interoperability issues, and risks repeating costly mistakes. Industry is often best-positioned to spearhead drafting of technical standards, as AI systems evolve rapidly, and standards must remain flexible enough to incorporate state-of-the-art research. Given that the governance of frontier AI systems involves highly specialised safety and security expertise, broad standards organisations may lack the bandwidth to focus purely on these more narrow risks. Consortia like the Frontier Model Forum have a mandate to focus exclusively on frontier AI issues, establishing frontier-specific best practices without impacting the development of more general AI models or applications.

The adoption of industry standards also serves as a crucial trust-building tool with potential customers. As organisations adopt AI, they may expect independent verification of security and governance controls to justify that trust. By conforming to recognised frameworks such as ISO/IEC 42001 and ISO/IEC 27001, developers of AI agents can ensure foundational governance and security practices are implemented.

National institutions like the U.S. Center for AI Standards and Innovation (CAISI) or the UK AI Security Institute also serve a vital catalytic role in convening industry, academics, civil society, and other stakeholders to develop best practices. These institutions are also key in stress-testing proposed

protocols, and championing frameworks in international standard-setting bodies, such as via efforts like CAISI's AI Agent Standards Initiative.

Much as the development of cybersecurity standards took decades, building out agent security standards will be a gradual, multi-year process. A crucial first step will be standardising definitions and core functional protocols to create a foundational layer for both interoperability and security. Open protocols like A2A, AP2, and UCP handle how agents interact with each other, payment services, and retail sites, respectively. These could become vital as the backbone of agentic commerce, much as open communications protocols like TCP/IP, identity protocols like DNS, and security protocols like HTTPS, OAuth, and TLS/SSL now form the backbone of the Internet. However, just as standardising TCP/IP did not eliminate the need for enterprise security architectures, establishing secure agentic communication protocols is likely to only be the beginning of this work.

Recommendations: pre-standardisation work to advance agent security

Because society is just beginning to develop and deploy AI agents, the policy frameworks we adopt must remain agile enough to ensure that regulation remains effective and does not inadvertently stifle the innovation that drives technological breakthroughs. The challenges of agentic AI often do not require policymakers to reinvent fundamental goals and frameworks. As Google's TRUST Framework outlines, we need quick and nimble industry standards around governance, oversight, safety, and security to manage agentic capabilities and potential risks. The rise of AI agents calls for reaffirmation of responsible practices around AI, and stakeholders must work together to create the clear, consistent guidelines needed for the public to trust and adopt AI agents, and to help address emerging agentic risks.

As guidelines for agents are being developed, they should take into account existing cybersecurity principles and frameworks, which are battle-tested and provide a strong starting point. For example, Google DeepMind's AI Control Roadmap treats certain AI agents as untrusted, drawing on traditional cybersecurity strategies like "zero trust," which remains a powerful north star for managing AI agents. Industry and governments must work together to develop best practices and guidelines that balance trade-offs between security and usability, while working towards formal international standards that are ready for widespread adoption.

Two avenues of pre-standardisation work can help lay broader foundations for secure and widespread AI adoption:

- **Consensus-building:** Better clarity on desired objectives can make trade-offs clearer, and improve technical design choices. It can also improve coordination across different standards developers, avoiding a fragmented landscape of overlapping frameworks. AI Security Institutes (AISIs), like CAISI and UK AISI, can facilitate this consultative process, such as through CAISI's January 2026 Request for Information on AI Agent Security.

- **R&D:** Without adequate exploration of the design space for agent security, we risk prematurely locking in suboptimal choices for standardisation. Policymakers can facilitate R&D by creating space for industry and academics to innovate, such as with regulatory sandboxes, and via funding for academic research from research agencies.

To ground these pre-standardisation efforts in practical applications, it helps to examine the integration of AI agents through the lens of traditional enterprise security. By exploring three illustrative domains, we can more concretely identify how cybersecurity principles apply to AI agents, and where consensus-building and R&D can adapt them for the agentic era: identity and access management (IAM), secure software development practices, monitoring and detection.

Identity and access management: Verifying who is accessing a system and what they are allowed to do is a foundational challenge in cybersecurity. The management of non-human identities (NHIs) in this space is already a well-trodden problem, with common estimates indicating that NHIs (e.g., for admin bots and service accounts) outnumber human identities by 25–50x in modern enterprises. In this sense, IAM for AI agents is not a wholly new challenge, but practitioners will need to navigate a shift toward more autonomous AI agents, an increase in the overall volume of NHIs, and possibly longer-lived AI agents, as shown by agent-to-agent social networks like Moltbook. As the interactions between websites, agents, and browsers evolve, IAM solutions may need to adapt to keep the web both open and secure.

One set of solutions is agent identifiers ("agent IDs"), which some researchers have suggested could help prevent fraud, trace security incidents, and distinguish trustworthy agents from malicious bots or compromised agents. Many industry actors already implement agent IDs in specific contexts: for example, Gemini Enterprise Agent Platform lets developers assign unique identities to their agents to authenticate and authorise agent actions to enhance security. A line of work that some researchers have proposed is making agent IDs more interoperable across selected platforms and organisations, as this could potentially have security benefits including improved detection of malicious or compromised AI agents.

However, it may be the case that not every agentic application benefits from agent IDs. Some agent instances may not need IDs, such as instances that are more ephemeral (i.e., existing for only seconds or minutes to carry out a task). Long-lived agents may have varying tasks and context over time, complicating whether an agent ID is useful and actionable for authorisation or monitoring.

Agent IDs may also be more relevant for some use cases than others, such as for certain finance and healthcare use cases where failure could have a larger impact. As with other security interventions, usability and privacy must also be weighed when determining whether to implement agent ID. For example, agent IDs and the data associated with them could be sensitive from a commercial and legal standpoint, meaning privacy must be carefully considered when implementing any approaches to share or aggregate agent ID-linked data.

To investigate this line of work further, researchers in industry and academia could conduct field trials to pilot the use of agent IDs for specific purposes, such as detecting or disrupting compromised AI agents. As they do so, they could consider partnering with specific sectors to more concretely explore implementation benefits and challenges in real-world settings. Researchers could also conduct R&D on privacy-preserving technology to support agent ID infrastructure, making it easier for organisations to reap the security benefits of greater interoperability while continuing to protect sensitive user and organisational data.

As these field trials and policy conversations evolve, stakeholders should aim to establish clear objectives for well-defined agent ID schemes, rather than allowing technical infrastructure to precede the purpose for which it is built. By clearly defining the objectives for specific agent ID schemes, coordinating between different stakeholders' objectives, and thoughtfully weighing the implementation trade-offs between security and usability/privacy, the industry can more confidently move to advance IAM standards that are fit-for-purpose.

Secure software development practices: To help downstream developers better secure their agentic applications, industry can build on current software development standards by laying out related best practices for agents. Existing pre-standards like NIST's Secure Software Development Framework (SSDF) act as a widely used baseline for general software applications, and other pre-standards like NIST SP 800-218A have sought to extend these practices to foundation models. Agents take this one step further, as they combine an underlying foundation model with applications that use scaffolding, memory, and external tools.

Industry best practices that emphasise a secure-by-design approach can help create a consistent baseline for security against efforts by bad actors trying to subvert or redirect agents' activities. Google's Secure AI Framework addresses controls throughout the AI development lifecycle, covering risks like sensitive data disclosure and rogue actions. The security principles laid out in CoSAI Principles for Secure-by-Design Agentic Systems guide Google's internal implementations, and are intended to help users and industry more broadly.

Secure development frameworks will likely be most helpful if tailored to different audiences based on their needs and the risk profiles of their applications. For developers building basic workflows, a simple "Agentic Hygiene Checklist" could minimise immediate risks, ensuring that developers do not neglect common practice, such as following the principle of least privilege. As enterprise applications continue to evolve, the industry can consider more comprehensive frameworks as needed, such as for higher-risk enterprise applications.

To lay the groundwork for such higher-risk agentic applications, industry can conduct further R&D to develop state-of-the-art defences and refine them for use by downstream developers in production environments. For example, Google has conducted research on contextual agent security to generate just-in-time, contextual, and human-verifiable security policies. Google DeepMind also developed a Capabilities for Machine Learning (CaMeL) defence, where architectural isolation separates trusted task planning by a user from untrusted environment observations and potential prompt injections.

Ultimately, to ensure the adoption of best practices around agent security, governments can also play a part by making security a core part of procurement. Doing so can encourage vendors to build robust security features into commercial products and raise the security baseline across the market, facilitating a more secure ecosystem overall.

Monitoring and detection: As AI agents become more autonomous, the use of AI agents to provide automated scalable monitoring ("AI monitors") of other AI agents is emerging. Because of the scale and speed at which AI agents operate, traditional mitigations relying on human oversight can become costly

or insufficient. The use of AI monitors can be a way to complement other more established measures for detecting undesired AI agent behaviours. AI monitors are not yet battle-tested, meaning further R&D and field trials are needed to identify how accurate, cost-efficient, and reliable they are. Determining when, where, and how such AI monitors can be implemented and tested is an open research question, where industry could develop best practices.

To explore these research questions further, Google DeepMind is beginning to experiment with the use of AI monitors in certain deployment settings. For example, its AI co-clinician uses a dual-agent architecture where a "Planner" module monitors conversation continuously to verify that the "Talker" agent stays within safe clinical boundaries. For future more powerful agentic AI systems, Google DeepMind's Approach to Technical AGI Safety describes the use of "AI monitors" to detect potentially dangerous actions, while the AI Control Roadmap describes a monitoring approach for internally deployment AI agents, an ideal testbed given Google's leadership in frontier AI development and deep expertise in cybersecurity.

Part II: Addressing multi-agent risks

Part II: Addressing multi-agent risks

As agentic capabilities scale, ecosystems where multiple AI agents interact, negotiate, and collaborate will become commonplace. Some sectors have already had a taste of this: in finance, for example, high-frequency trading algorithms have been interacting with each other at scale for decades. But modern AI agents could represent a step change, with their greater autonomy and context-awareness enabling a larger and richer suite of interactions in a more diverse range of fields. Decentralised multi-agent setups are inevitable in an interconnected world with single-agent adoption. At their best, high-fidelity AI agent proxies could even amplify human agency by enabling "Coasean bargaining at scale," where agents acting as our faithful representatives in negotiating the frictions of everyday life could reduce negotiation costs to near-zero, turning previously unsolvable zero-sum dynamics into newly feasible, mutually beneficial arrangements.

While multi-agent systems have been a robust field of study for decades, we are still learning about how modern, LLM-based multimodal models change multi-agent dynamics. Just as many humans interacting can create efficient markets and produce novel inventions, ecosystems constructed from such agents could streamline logistics coordination and accelerate scientific innovation. Yet even basic questions, like under what circumstances multi-agent systems outperform individual agents, are still the subject of active research. In a recent paper, Google researchers challenged a common assumption that "more agents are always better," finding that with current AI systems, multi-agent coordination is most useful when applied to parallelizable tasks, but often counterproductive for sequential ones.[2] However, this may not necessarily be true of future AI systems if delegation mechanisms become more developed.

Multi-agent failure modes

Multi-agent systems also carry risks that do not manifest when agents operate in isolation. As argued in Google Deepmind's work on Distributional AGI Safety, perhaps most critical is the risk that AGI-level capabilities may first emerge not within any single model but across a network of specialised sub-AGI agents with complementary skills and tool access—a phenomenon the authors call "patchwork AGI." Because no constituent agent is individually an AGI system, this emergence could go undetected by safety frameworks designed to evaluate such systems. Moreover, as multi-agent interactions may produce adverse outcomes even if individual AI agents are aligned with user intent, one concern is that efforts focused on individual agent-level alignment may fail to provide adequate safeguards at the level of a multi-agent AGI system.

Even before that threshold, agentic collectives may already introduce distinct failure modes. In a smart grid, cascading errors could theoretically cause blackouts; in a financial system, miscalibrated agents could trigger bidding wars. Multi-agent systems could produce undesired behaviours, like tacit collusion where agents learn to coordinate harmful strategies without explicit communication, or vulnerabilities stemming from algorithmic monocultures across similarly architected models.

Single-agent frameworks can fall short in addressing these problems. As an example from Google DeepMind's work on Intelligent AI Delegation, many multi-agent systems rest on delegation, where

[2] This is true for current AI systems in part because the lack of intelligent delegation can make it difficult to ensure robustness and extract value from longer sequential agentic chains; parallelization can be seen as the best "simple" solution in the absence of strong delegation frameworks.

an agent breaks complex goals down into sub-tasks, then routes them to specialist agents. However, improper delegation can produce accountability vacuums where harm occurs, but no single node is clearly responsible. While overtly malicious requests are weeded out by single-agent safety filters, most other directives fall into a "zone of indifference" where downstream agents dutifully execute tasks without critical scrutiny—even if they do not fit the original intent or broader context. Current delegation chains are typically short, but as agent adoption increases, they will grow in length, meaning that such subtle mismatches may compound more rapidly, and with greater consequences.

Weaponisation of the information environment

Furthermore, the information environment itself can be weaponised against AI agents, adding another layer to these arising risks. As described in Google DeepMind's research on AI Agent Traps, attackers can embed adversarial content in the shared resources agents consume. Systemic traps could allow attackers to target and exploit weaknesses in collective system dynamics:

- **Congestion traps** could synchronise homogeneous agents into exhaustive demand for limited resources, triggering self-inflicted denial-of-service;
- **Interdependence cascades** could amplify a single fabricated signal through reactive agent networks, akin to the 2010 flash crash in the stock market; and
- **Compositional fragment traps** could partition a malicious payload into benign-looking pieces distributed across multiple data sources, which only reconstitute into a full attack when agents aggregate them.

AI agents' potential susceptibility to mass compromise via these shared environmental attack surfaces could create broader risks—for example, a single poisoned financial report could trigger a synchronised sell-off among thousands of agents sharing similar architectures and reward functions. Beyond their economic impact, these systemic traps could also have a national security impact: for example, as the public increasingly relies on agents and agent-produced content, such traps could make it easier for attackers to conduct large-scale campaigns of harmful manipulation.

This homogeneity, and the interconnectedness and complexity of agent-to-agent interactions, might make such large-scale attacks difficult to trace and counter. Participants in human markets can be "hacked" via coercion, bribery, and honeypots, but rarely are such attacks feasible at large scale or conducted synchronously. By comparison, attackers could use weak points in inter-agent communication to tamper with AI agents' messages and decisions at scale, or adopt novel techniques like "Morris 2.0," a self-spreading prompt injection, to infect and subvert large numbers of AI agents.

Recommendations: interdisciplinary research to address multi-agent risks

To address potential problems such as these, multi-agent safety can benefit from a layered, systems governance approach. Google DeepMind researchers have argued that multi-agent systems may begin to increasingly function as "virtual agent economies" as AI agents are more widely deployed, and begin to interact in more complex ways in the real world. This suggests that frameworks for managing multi-agent risks could incorporate not just technical measures but also draw on market-based governance approaches.

Interdisciplinary research straddling AI, cybersecurity, economics, and other fields could boost our confidence in how to govern a world with many interacting AI agents. Key research avenues include understanding multi-agent risks, piloting market infrastructure, and experimenting with market design.

For example, Google DeepMind, together with Schmidt Sciences, the Cooperative AI Foundation, and the Advanced Research and Invention Agency have announced a technical research funding call of up to $10M to support the study of how large-scale multi-agent AI systems behave as a group, and how we can provide frameworks to understand and mitigate against potential risks.

Understanding multi-agent risks: Before committing to specific policy mechanisms that address multi-agent risks, policymakers might want to invest in R&D, as well as consensus-building to align on which multi-agent risks might deserve the most attention. National AISIs (e.g., the UK AISI or US CAISI) and researchers in industry and academia could conduct further empirical research into concrete multi-agent threat models, building off existing work that has identified potential high-level risks like cascading failures, conflict escalation, and tacit algorithmic collusion.

Such research could address both organically emerging anomalies and large-scale adversarial attacks, which may overlap: for instance, poisoned databases could trigger synchronised sell-offs, or contagious prompt injections could trigger collusive behaviour. Researchers may want to consider not just approaching these issues from a purely theoretical perspective, but tapping on existing empirical data to inform their work. For example, existing cybersecurity technologies (such as Censys, Shodan, or Greynoise) can provide valuable insights about the nature and behaviour of various digital assets and their behaviours at scale on the Internet.

Research on multi-agent risks could also tap on domains like economics to better characterise multi-agent dynamics. Google DeepMind research on virtual agent economies has identified these emerging virtual economies as varying in their permeability—that is, the degree to which they are interconnected with or buffered from real-world economic systems. As the use of AI agents grows, it will be helpful to understand how these buffers vary across different sectors, and whether they are appropriate to the evolving multi-agent world.

Building market infrastructure: As these virtual agent economies develop, industry and academic researchers could also explore what architectural elements would help to provide a stable foundation for multi-agent ecosystems. For example, some researchers have argued that establishing certain rules of participation for AI agents could potentially help to embed accountability into market operations and shield the real-world economy from agent-driven volatility.

One hypothetical example of such market infrastructure is delegation protocols to ensure accountable relationships between market participants (i.e., AI agents). Suggested by Google DeepMind researchers, the idea is that such protocols could, in theory, help prevent future accountability vacuums as agent-to-agent chains of interaction grow in complexity. Such protocols would treat delegation as more than just task assignment: similar to human chains-of-command, they would require the explicit transfer of authority, responsibility, and accountability.

Industry actors could consider adapting protocols like MCP and A2A to explicitly include new data fields

that implement "intelligent delegation." As an example of one such field, verification standards could allow two agents to agree on what a successfully completed task looks like before handoff. As these protocols evolve, they could include more complex requirements, such as "firebreaks" where an agent must decide whether to assume responsibility for an output or halt execution and escalate to a human, thus preventing unchecked errors from cascading down the delegation chain.

Experimenting with market design: Establishing foundational market infrastructure might, in the longer term, then let us use market design as a lever to mitigate harmful collective behaviours. Google DeepMind research describes the use of a trust and reputation architecture to influence how market participants like AI agents behave. In theory, this could permit a more scalable approach to governing multi-agent ecosystems. Low-trust agents might face more constraints and lower earning potentials, potentially creating economic incentives for agents to act reliably without requiring exhaustive human oversight.

However, such an architecture assumes the widespread use of long-lived AI agents, and its exact details would rest on how such agents function and what other infrastructure they use. As such, a trust and reputation architecture, like other market design elements, may sit atop foundational elements like agent identifiers and standardised metrics for task success.

Developing implementation-ready market designs will likely require significant R&D. As the underlying elements continue to develop, researchers in industry and academia could experiment with proof-of-concept implementations, such as behavioural metrics (e.g., safety scores) and tamper-proof performance ledgers to record task outcomes. In designing these, they could look not just at traditional computer science, but also toward fields like economics, finance, and sociology, where multi-agent dynamics are already salient.

To catalyse this cross-disciplinary research, national science and research foundations, along with private philanthropies, could sponsor grand challenges to address problems in this space, such as algorithmic market stability. These grand challenges could provide academic researchers with the funding and compute to model market design systems in high-fidelity agentic simulations. Eventually, policymakers could consider advancing standards, infrastructure, and pilot tests to enable these market designs to move from theory to reality. They could do so, for example, by developing regulatory sandboxes to stress-test these designs on a limited scale before expanding them to the larger economy.

Part III: Empowering cyber defenders and building resilience

Part III: Empowering cyber defenders and building resilience

The availability of AI agents also has wider impacts for both cyber defence and offence. As AI agents become more capable, defensive agents could transform cybersecurity for the better by bolstering detection and response, as well as eliminating software vulnerabilities. However, this outcome is not a given: it will require thoughtful investment in ensuring at-risk defenders can make systematic use of these tools, and in ensuring AI agents are securely deployed and implemented. As attackers themselves make increasing use of AI agents, the pressure will be on defenders to adopt defensive AI agents safely and securely; policymakers should work with them to navigate this transition.

Facing an accelerating operational tempo

The accelerating operational tempo in cybersecurity means that technology companies and policymakers have to begin thinking about how to harden digital systems and accelerate defensive security operations now. Most immediately, AI systems have the potential to uplift attackers, increasing the speed, scale, and sophistication of cyberattacks that they can launch. As bad actors adopt AI, defenders are experiencing increased pressure, especially under-resourced organisations like operators of lifeline infrastructure. These actors have traditionally struggled to implement security best practices, let alone adopt new technologies, and policymakers should be careful not to leave them behind.

In the longer term, more capable AI agents could fully automate complex malicious cyber operations. Such "highly autonomous cyber-capable agents," if developed, might be akin to intelligent botnets and use more adaptive techniques to evade traditional defences, such as "agentic implants" that operate locally on victim networks so that callbacks to their operators do not give them away.[3] These sophisticated agents could also pursue new targets, such as computing resources, which they may need to sustain their own operations, or cyber-physical systems like household and industrial robots that may be vulnerable to attack as their real-world use rapidly grows.

The promise of agents for cyber defence

AI agents could help defenders improve their security posture in at least two ways: firstly, by ensuring a secure software development lifecycle and second, by accelerating the work of cybersecurity practitioners. Much of our digital infrastructure's insecurity stems from vulnerable software, and AI could help automate the detection and remediation of these vulnerabilities. Google has built agents like CodeMender to autonomously develop patches and use them to improve the security of open-source projects. These tools need to be in the hands of defenders, and Google is working with the rest of industry to accomplish this, as with a recent joint pledge of $12.5 million to invest in the stability and security of the open-source community. To counter malicious intrusions, agentic security operations centers (SOCs) can also help cyber practitioners more quickly process vast volumes of security logs and free up time for more complex investigations—a crucial aid, given that SOC operators for years have consistently reported high levels of burnout.

Using agents to implement secure software at scale could prove especially transformative. This could

[3] Traditional malware implants require persistent communication traffic between the victim's and the attacker's computer to receive instructions from human operators, which increases their detectability.

vastly reduce the attack surface that defenders need to cover. Indeed, the potential for AI to resolve systemic vulnerabilities has led some specialists to suggest that the technology could eventually signal "the end of cybersecurity" as we know it, by eliminating large swaths of vulnerabilities before attackers have the chance to exploit them. Some experts have spoken about how AI could enable moonshot goals like rewriting millions of lines of insecure code, or permitting the automated formal verification of software to prove that systems are hardened against entire classes of attacks.

The need to secure cyber defence agents

As AI agents bolster cyber defences, the delegation of responsibility to such agents will make them a prime target for attack. Subverting the AI systems in charge of an organisation's network defences could allow attackers to gain widespread access while remaining undetected; poisoning the training data of a "secure" coding agent might eventually let attackers insert backdoors in codebases that are assumed to be safe. This makes it especially important to build out the field of agent security, and to ensure that AI agents are not given full trust where it is not justified. Instead, organisations deploying AI agents for cyber defence should apply a zero-trust approach, as discussed earlier in this paper, by implementing checks (e.g., monitoring by other AI agents) to detect and monitor for unintended behaviour and compromise.

As third-party security and coding agents become more common, it will also be increasingly important to establish standards for trust in the agent supply chain. The rapid growth of community-contributed agent capabilities highlights how easily threat actors can exploit unvetted AI ecosystems. For instance, the OpenClaw marketplace was recently targeted by the ClawHavoc campaign to distribute hundreds of malicious agent skills. This can be seen as the AI-era evolution of supply chain exploits like the XZ Utils attack, where malicious actors gained the trust of an overworked open-source maintainer for three years before inserting a backdoor. Because these are attacks on marketplace trust rather than on underlying agent security, the efforts required to address them needs to be structural. Preventing malicious agent distribution may require a trust and reputation architecture for third-party skills and coding agents, in addition to the zero-trust measures described above.

Recommendations: a coordinated effort to improve cybersecurity

To ensure that society can meet the cybersecurity risks presented by frontier AI capabilities, defenders will need to invest in a wide range of efforts. Policymakers, industry, and academia should collaborate to advance initiatives across five main goals:

First, strengthen understanding of the threat landscape by monitoring progress in AI capabilities and adoption by threat actors. Second, secure software products and services by (a) fostering a secure-by-design approach and (b) fortifying supply chain integrity. Third, for digital systems in deployment, (a) reduce the attack surface and (b) adopt AI agents for security operations. Fourth, empower and mobilise defenders by (a) investing in defensive AI tools, and (b) bolstering talent and institutional capacity. Fifth, improve collective visibility into real-time threats.

1. Monitor the progress and diffusion of AI-enabled cyber capabilities

To strengthen our understanding of where the cyber frontier is and how AI may transform the threat landscape, AISIs, frontier AI developers, and academics could invest further in cyber capability evaluations. These evaluations can include not just a focus on vulnerability discovery or capture-the-

flag tests, but also realistic multi-stage attack scenarios. This can include cyber ranges that mirror critical infrastructure systems and networks, such as UK AISI's "The Last Ones" and "Cooling Tower" simulations. As frontier AI systems improve, future testing could consider using defensive AI systems as part of the evaluation, with such "red-on-blue" simulations providing more insight into how frontier AI systems perform in real-world environments when evading defences.

Beyond direct investment, these parties can also invest and coordinate on improving the science of AI evaluations and underlying evaluation infrastructure, as bad evaluations can be potentially misleading. Where possible, we should aim to achieve international alignment around standard testing practices, such as via coordination through the International Network for Advanced AI Measurement, Evaluation, and Science, so as to avoid duplicative testing.

However, evaluations can only tell us about the current state and past development of AI cyber capabilities. As such, intelligence and cyber agencies, as well as academics, can also conduct forecasting exercises and monitor diffusion dynamics to model how quickly different threat actors will adopt advanced AI-enabled cyber capabilities, as this may change which sectors or targets need to be hardened first. For instance, national intelligence agencies could designate adversarial use of agentic AI as a priority for intelligence collection, enabling governments and industry to make more informed decisions about threats to critical infrastructure.

2A. Foster a secure-by-design approach by improving software product security

Because AI adoption will increase the speed of malicious attacks, ensuring that software products—and AI tools—are secure from the start will become even more important. This means that vendors can play an important role in mitigating security weaknesses. Google has signed CISA's Secure by Design Pledge, and its 2024 commitment describes the steps toward pledge goals like adopting multi-factor authentication for users, avoiding default passwords, and reducing classes of common vulnerabilities (e.g., SQL injections, or memory safety vulnerabilities).

Governments should encourage vendors to develop secure-by-design products. Cyber agencies should continue to engage industry partners, such as the Secure-by-Design guidance published in a joint international collaboration in 2023. Agencies responsible for procurement and digitalization, like the U.S.'s OMB and GSA, the UK's DSIT, or Japan's Digital Transformation Agency, should also prioritise systems and products that are secure-by-design. One class of products to highlight here are trusted security devices. For instance, Mandiant's M-Trends 2025 report noted that the four most frequently exploited vulnerabilities were all in edge devices like VPNs, firewalls, and routers, a trend that is accelerating.

As AI transforms the discipline of software engineering, developers of AI coding tools, academic researchers, and policymakers can also work together to improve the security of AI-generated code. This could help produce secure code at scale, rather than producing insecure code that creates new attack surfaces, as some studies have warned about. Google DeepMind has invested in creating tools like CodeMender, which can support developers by autonomously reviewing existing code and creating patches, and is conducting further research on how to improve the security of AI-generated code.

Academic and industry researchers can conduct research into this area and related benchmarks to reveal and boost novel strategies for reducing code insecurity, such as SecureForge, an automated

pipeline that optimises AI prompts to improve the security of outputted code. R&D agencies, like DARPA, UK ARIA, or the ERC, can consider investing in standardised security benchmarks for secure code verification and generation to incentivise a race to the top among developers of AI coding tools on code security.

2B. Protect software supply chain integrity by raising the bar on developer practices
As AI uplifts bad actors and the software ecosystem becomes more complex, we should also be prepared to see more software supply chain attacks (i.e., nefarious alteration of trusted software before delivery), as shown by a spate of such attacks including TeamPCP's March 2026 attacks on the Trivy and then LiteLLM software libraries. This is a particularly concerning vector for attack because malicious actors could successively steal large numbers of developer credentials, enabling them to conduct "snowball" attacks that can reach a huge blast radius: LiteLLM, for example, is an open source package with over 90 million users.

Rather than requiring software vendors to improve the security of their products (e.g., by addressing code vulnerabilities), software supply chain integrity requires software developers to change their own personal and organisational practices to prevent, for instance, their developer accounts from being hacked and used to distribute malware. Better supply chain integrity can be facilitated by a broad-based effort that aims to establish common standards and frameworks to help everyone raise the bar, as well as more targeted efforts to equip the open-source community with the tools and resources to implement these frameworks.

In Practice: How Google has supported supply chain security

Google has invested significant effort in developing frameworks to help actors across the ecosystem secure the software supply chain:

- **SDLC infrastructure security:** To prevent breaches from happening in the first place, upstream developers need to protect their software development lifecycle (SDLC) infrastructure. The Wiz SDLC Infrastructure Threat Framework (SITF) helps developersthem model malicious attacks (e.g., attackers compromising developer accounts) and put the right controls in place.
- **Software provenance:** Upstream developers also must be able to prove to downstream consumers and developers that their code can be trusted. The Supply-chain Levels for Software Artifacts (SLSA) framework, which Google developed and open-sourced, helps them generate attestations that prove that software releases have not been tampered with.

Google has also worked closely with package managers and open-source developers to secure their software. For example, Google co-developed and open-sourced Sigstore to help developers validate the integrity of software artifacts, and in 2024, helped integrate it into the Python Package Index (PyPI), one of the largest open-source package registries in the world. This makes it easier for hundreds of thousands of project maintainers to prove to downstream users that they are downloading legitimate software updates.

A broad-based effort to establish common standards and frameworks for supply chain integrity could help everyone raise the bar, ensuring tamper-resistance across the pathway from source to consumer. To improve the security of software used by the government, digitalization and procurement agencies can take into account existing supply chain security frameworks when defining cybersecurity requirements for vendors. This will help prevent upstream software packages from being compromised, although downstream developers may also need to take their own measures to defend against importing bad packages, such as not automatically downloading and updating new packages.

At the same time, more targeted efforts are needed to equip the open-source community with the tools and resources to implement these frameworks. Open-source maintainers have limited resources and underpin much of the software ecosystem, making them attractive targets, as shown by the TeamPCP attacks. Working with central platforms (i.e., package registries like PyPI and npm) to incorporate secure SDLC practices is a high-leverage way to improve supply chain security practices in the open-source community, as it relieves individual open-source developers of the burden of figuring out what practices they need to adopt.

To further bolster supply chain integrity in the open-source ecosystem, government agencies can support key actors, like the software foundations that operate package registries, to improve their security posture. National cyber agencies could provide funding to open-source foundations and maintainers to expand their capacity, while research agencies could develop tooling that makes it easier for them to adopt supply chain best practices. For example, the NSF's Pathways to Enable Secure Open-Source Ecosystems could provide one such avenue to develop more helpful tooling for the open-source community.

3A. Have organisations proactively harden their deployed systems via posture management and infrastructure hardening

Ensuring system security through secure deployment is just as important as code security. Most cyberattacks are possible not because of code security (i.e., vulnerabilities), but due to weaknesses in network configurations or human judgement. Modern, cloud-based software and infrastructure can often help reduce the chance that organisations fall victim to such errors, compared to older legacy infrastructure.

On the other hand, there is also a long tail of un-maintained or poorly-maintained software for which improving code security is not possible. For example, for some Internet of Things devices, the original company may have gone out of business with the source code all but lost; or for critical infrastructure environments, there may be cases where patching and taking the system offline introduces unacceptable risks. Such scenarios can be particularly difficult to navigate from a cyber policy perspective.

Identify weaknesses, and fix what you can: As attackers use automation for reconnaissance and exploitation, enterprises and governments should increasingly look toward automation to improve visibility and remediation time. AI can be a force multiplier across the whole chain: organisations can use automated asset discovery and red teaming to find weaknesses, and automated remediation and patch-management tooling to close them faster. As examples:

- Wiz uses AI agents to proactively reduce the attack surface: a "red agent" scans the attack surface and uses contextual information (e.g., cloud, workload, and code analysis) to discover immediately exploitable risks, while a "green agent" helps customers identify root causes and automatically deploy fixes using pre-built Wiz skills.

- Offensive security startups like XBOW and Dreadnode are also offering automated red teaming, which can make deep system security engagements accessible to a broader audience. Such services could be helpful for vulnerable actors in areas like healthcare, telecommunications, and local government.

Harden, and replace what you can't fix: Where systems can no longer be fixed, or where patching can introduce safety and uptime issues, organisations may face more complex trade-offs. However, neglecting these systems can create dangerous weak points.

- Organisations should aim to deprecate and replace end-of-life systems where possible. These legacy systems (e.g., old networking appliances or operating systems) can be hard or impossible to patch and are at greater risk than modern, cloud-based software and infrastructure. Procurement and digitalisation agencies should strongly encourage government agencies to deprecate and replace end-of-life systems, and national cyber agencies should do likewise with critical infrastructure owners and operators.

- Research agencies and philanthropic funders can also use pull mechanisms (e.g., competitions like DARPA's AI Cyber Challenge for patching vulnerabilities) to incentivise the development of AI-enabled tools that enable automated hardening, and compensating controls (such as anomaly detection or airgapping) when patching is not possible. While such tools should not be seen as a substitute for replacing end-of-life systems, they may be especially valuable in helping under-resourced and critical organisations navigate the constraints that they face.

3B. Adopt AI agents for security operations to scale network defence against AI-enabled threats
Promoting agentic AI to securely scale network defence and disruption will also be crucial to ensure that defenders have an advantage relative to adversaries. At Google, the Agentic Security Operations Center, powered by multiple connected and use-case driven agents, can execute semi-autonomous and autonomous security operations workflows on behalf of defenders. By leveraging agents for activities such as performing initial triage, gathering forensic data, or executing containment protocols, organisations can significantly reduce their threat detection and response time. Similarly, startups like Asymmetric Security are developing agentic systems to fully automate digital forensics and incident response (DFIR), transforming manual cyber investigations into a scalable process that can take hours instead of days.

Governments and industry can collaborate to deploy AI tools to harden the security of government infrastructure and critical infrastructure before powerful AI-assisted offensive capabilities spread. To do so, governments can fast-track supplier onboarding and provide federal funding for adoption of agentic security tools, with a focus on automated response and detection capabilities. In some regulated sectors, like finance, government legislatures can consider creating regulatory sandboxes to facilitate experimentation with advanced cyber capabilities, enabling organisations to test these tools under

regulatory supervision without fear of penalties due to compliance requirements.

As we do so, we should also aim to preserve trust in these tools. From SolarWinds to the recent Trivy breach, history shows that cybersecurity products often become targets, because such tools are often highly trusted and can become a valuable backdoor for skilled attackers. Frontier AI developers, cybersecurity companies, and research agencies should conduct more research into how to guard defensive cyber agents against exploitation. This could include safeguards against cyber defence agents being fooled by adversarial attacks into reporting false negatives, or secure code generation tools being poisoned to insert backdoors. It could also draw on work on AI monitors to detect rogue actions. Such work could lay the foundation for future standards and best practices around developing and deploying defensive cyber agents, ensuring that we can reliably trust these tools to protect us from malicious AI-enabled threats.

4A. Invest in defensive AI tools to build out a robust agentic security ecosystem

The cyber ecosystem benefits from having a diverse stack of AI-cyber capabilities usable at a wide range of price points. Defenders vary in terms of their security budget and their technical needs: one team might experiment extensively with powerful and costly frontier models to maximise performance, while another may want a suite of specialised, cost-effective tools to help them improve their cybersecurity posture.

Frontier AI developers can lay the foundation for this diverse stack by building a broad spectrum of models, ranging from highly capable frontier models to accessible, open-weight models. Third parties and frontier developers, however, will still need to build scaffolding on top of foundation models. This can elicit improved performance, particularly for specialised use cases, and turn models into more broadly usable products. Google's CodeMender, for example, helps automate not just finding, but fixing, of vulnerabilities, while adding an automatic validation process that ensures it only surfaces high-quality patches for review.

To ensure that defenders have the right toolkit for their needs, policymakers can play a role in identifying areas that the commercial market might neglect and spurring innovation in them. For example, some researchers have suggested that since critical infrastructure operators have to balance patching against safety and uptime, they might benefit from specialised LLM-based tools tailored to their capacity and needs, such as:

- Patch triage and automated network analysis to ensure that critical patches are prioritised for application;
- Autonomous hardening tools to put in place compensating controls when patching is not possible; and
- Defensive tooling to monitor if unpatched vulnerabilities are being exploited.

The innovation space is vast, and research and cyber agencies should work with civil society and academics to identify what gaps exist, and what capabilities might be truly transformational in addressing them. As this becomes clearer, research agencies, private philanthropies, and other actors provide monetary and other incentives to startups and open-source developers to "design for the defenders who need it most."

Goal 4B. Bolster talent and institutional capacity for defenders across the ecosystem
To realise the benefits of AI-enabled tools, many under-resourced entities will need more than just access and tools. They may also need talent, funding, and institutional capacity to enable them to make the best use of these tools and to adopt other, existing best practices.

To improve the institutional capacity of public sector entities, the government can provide surge funding, training, and human capital. Agencies can fast-track supplier onboarding and provide federal funding for adoption of agentic security tools. National and regional cyber agencies, like ONCD, CISA, the NCSC, or ENISA, can also play a role in rallying agencies to act, such as by convening a 90-day government-wide sprint on vulnerability preparedness and national resilience.

To improve the institutional capacity of private sector actors such as critical infrastructure owners and operators, policymakers can also collaborate with other actors to design public service programmes that hire and train national teams of cyber experts to improve the security posture of vulnerable and critical organisations. Google has worked with the Consortium of Cybersecurity Clinics to support higher education institutions in building a workforce with the real-world experience needed to protect critical U.S. infrastructure from cyberattacks. National cyber agencies could also drive larger versions of such programmes, such as a "Security for America" programme in the U.S., in the spirit of TechCongress and AmeriCorps. These experts could provide part-time support to establish quick security wins, as well as providing coaching to ensure a sustainable security baseline moving forward.

Lastly, government cyber agencies and private philanthropic actors can expand capacity for coordinated vulnerability disclosure, such as by working with Carnegie Mellon's CERT Coordination Center to expand staffing and equip them with compute and state of the art AI tools. Today's AI tools are likely to produce a surge in newly discovered vulnerabilities that the current coordinated vulnerability disclosure system will struggle to manage. Boosting this via partnership with institutions like the CERT Coordination Center can ensure effective triage and coordination at a time when it is needed most.

Goal 5. Improve collective visibility into real-time threats by bolstering detection and information sharing of new, AI-enabled threats
Malicious actors are increasingly conducting campaigns across multiple AI platforms, attempting to evade detection by breaking up these campaigns into many smaller tasks that individually appear innocuous. To combat this, frontier AI companies, cloud providers, and other infrastructure companies (e.g., telecommunications and payments providers) can work together to explore shared security signals that enable detection of such campaigns while preserving user privacy.

While cross-industry discussion is still needed to determine the best approach for this, one way to lay the groundwork for greater threat intelligence sharing could be establishing secure forums to voluntarily exchange information, drawing inspiration from the Global Signal Exchange that Google set up in 2024 to combat fraud and scams. Such forums could allow organisations to track patterns of model misuse or malicious agent activity, and expose coordinated campaigns that would be invisible to any single organisation. Such a forum could take the form of an independent "Agentic Security Exchange," or otherwise be established as one function of another institution like the Frontier Model Forum, which has already established some functions for information-sharing among member frontier AI companies.

In the longer term, such cooperation could ultimately look toward more ambitious programmes or frameworks to take down and disrupt coordinated agentic campaigns, analogous to Google Threat Intelligence Group's recent work to take down the IPIDEA residential proxy botnet. The nature of this coordination will depend on technical feasibility and the threats we observe, which are still coming into focus.

Conclusion

Conclusion

As AI technology approaches the next point in its evolution, we—as technologists and policymakers—have agency in deciding whether the rise of agents will benefit cyber defenders or introduce new risks. Realising the promise of AI agents requires deliberate, proactive work to get things right early on, rather than waiting for vulnerabilities to spread.

This moment in AI development mirrors the early days of the Internet: a foundational period where actors were just starting to agree on the protocols, safety standards, and operational norms to govern a globally interconnected ecosystem. Because agentic systems will inevitably interact across complex multi-agent frameworks, the security of this new digital infrastructure must be embedded now, before these systems scale up. Securing this future is a collaborative endeavour that demands participation from across the digital landscape—model providers, downstream developers, policymakers, and civil society alike.

Acknowledgements

We would like to thank Kevin Klyman, Ellie Sakhaee, Sherry Huang, Jennifer Beroshi, Matija Franklin, Nenad Tomasev, Scott Coull, Sebastien Krier, Owen Larter, Allan Dafoe, Anna Wang, Tim Dierks, Lucy Lim, Celine Smith, Mary Phuong, Victoria Bew, Victor Escobedo, Erik Jenner, Lisa Einstein, Matthew Mittelsteadt, Jam Kraprayoon, Matthew van der Merwe, Asher Brass, Christopher Covino, Kamile Lukosiute, and Edan Maor for their reviews and valuable feedback. All views, and any mistakes, belong solely to the authors.