Google's approach to AI Agent Security

AI agents offer a monumental leap in technological capability and unprecedented productivity gains. However, their power to take action in the digital and physical world introduces critical and immediate security risks. Ensuring agents operate safely and reliably is a paramount concern demanding robust safeguards. Google is proactively addressing these high-stakes challenges with rigorous guiding principles and defense measures.

What are AI agents?

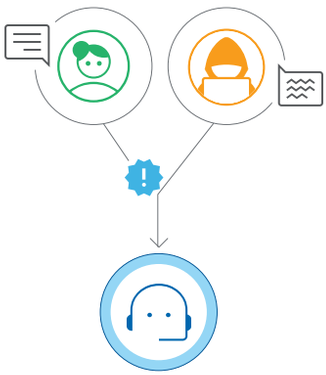
AI agents are AI-based systems designed to perceive their environment, make decisions, and take actions to achieve user-defined goals. Unlike standard Large Language Models (LLMs) that primarily generate content, agents interact directly with other systems to perform tasks. This capability spans simple automation, like categorizing incoming service requests, to complex planning like researching a topic across multiple sources, summarizing the findings, and sending team communications. It is this ability to *act* that necessitates an intense focus on security.

Why Agent Security is crucial

Securing agents involves a fundamental trade-off: the more independent, powerful, and therefore useful they are, the harder it is to ensure they don't take harmful actions. Traditional software security methods, like strict "yes/no" rules for specific actions, often lack the flexibility needed for adaptable agents. For example, a rule like "AI must never unlock the front door to my house" provides absolute security, but limits functionality for a home assistant. On the other hand, relying solely on the AI's own judgment is dangerously insufficient, because AI can be manipulated. This risk increases when agents have significant autonomy and access to high-risk actions, such as controlling medical devices or physical access systems. Balancing the benefits of independence with the potential for high-impact errors makes Agent Security an incredibly important and complex challenge.

Security risks associated with AI agents

AI agents can add significant value but can also introduce unique and critical security risks that require a dedicated focus. Two key threats stand out:



1. Rogue actions: When agents perform unintended, harmful, or policy-violating actions or do not follow the user's instructions.

- **Common cause:** "Indirect prompt injection," where malicious instructions hidden in processed data such as emails or documents hijack the agent.
- **Example:** A user asks an agent to summarize a document, but hidden instructions embedded in that document trick the agent into an unrelated action, like making an unauthorized purchase.

2. Sensitive data disclosure (data exfiltration): When agents improperly reveal private information.

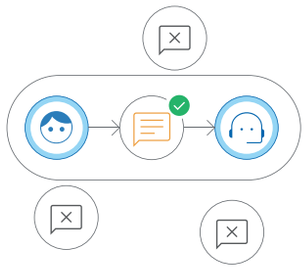
- **Common cause:** Attackers manipulate agents, often via "prompt injection" to retrieve and leak sensitive data.
- **Example:** Using prompt injection, an attacker might manipulate the agent into sharing a private document with the attacker.

Google

Google's principles for Agent Security

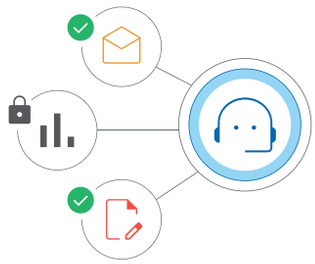
To address the risks associated with AI agents while preserving their utility, Google suggests three core principles:

1. Agents must have well-defined human controllers.



Agents can help humans with tasks such as sending emails or managing smart devices. For accountability, it's important to be able to trace key actions back to users. To avoid being tricked, AI agents must be able to clearly separate genuine commands from their users from any other instructions. They should also require explicit human approval before taking significant or irreversible actions, ensuring that the user remains in charge, with clear oversight over agent and user identity and permission.

2. Agent powers must have limitations.



Making an agent more capable can make it more useful, but also potentially more dangerous. That's why setting clear limits on an agent's abilities is essential to balance usefulness with security. By carefully defining what an agent can and cannot do—like allowing an email assistant to manage messages but not financial accounts—we reduce the risk of serious harm if the agent makes a mistake. This ensures agents have only the capabilities and permissions necessary for their intended purpose and cannot escalate their own permissions inappropriately.

3. Agent actions and planning must be observable.

To effectively manage and secure AI agents, users need to see what they are doing and understand their reasoning—like seeing the steps someone took to solve a math problem. This allows users to verify that the agent correctly followed instructions and achieved the right outcome, and helps them understand how the agent operates, building trust in the system. And when things do go wrong, this observability lets agent providers troubleshoot, prevent it from happening again, and spot suspicious activities, all consistent with privacy controls. As with traditional systems, observability allows for collecting information about a system's internal states and communication between its components, while keeping storage and access to user data secure.

Google's hybrid approach to mitigating AI agent risks

Google takes a hybrid approach to Agent Security. First, we make the underlying AI models more resilient through "adversarial training," which teaches the models to identify and resist prompt injection attacks—similar to how people learn to spot phishing scams. But just as humans can be tricked by a clever scam, AI models can also be tricked. That's why we add a second security layer through "policy enforcement." These checks review what the agent plans to do and compare the actions to the agent's security policies. Based on those policies, the action is either allowed, blocked, or the agent is prompted to ask the user for clarification ("Are you sure you want to spend more than $100?"). Policy enforcement acts as a crucial guardrail, enforcing boundaries that work alongside the agent's own judgement.

We are actively continuing to test, learn, and optimize our approach to Agent Security, recognizing the dynamic and evolving nature of this critical field. We are also collaborating with the "Coalition for Secure AI" (CoSAI) on their important workstream focused on "Secure Design Patterns for Agentic Systems." This collaborative effort underscores our belief that a shared understanding and collective action are essential to realizing the transformative potential of AI agents while effectively mitigating their inherent risks.

Google