

Agentic AI and its threats in the workplace

Artificial intelligence is no longer limited to writing text, summarizing files, or answering questions. Some systems have evolved into what is known as Agentic AI, capable of understanding goals, planning, utilizing tools, and taking practical steps to execute tasks fully autonomously to the user's behalf.

Examples of Agentic AI in action:



Meeting Scheduling:

User requests to schedule a team meeting for next week.

The Agentic AI searches for suitable times, selects a date/time, and prepares and sends the invitation based on the permission granted to it.



System Updates:

Granting the Agentic AI permissions on the system to periodically monitor the latest updates for software and the OS itself.

The agent then alerts the user to an available update and automatically updates the systems and software.

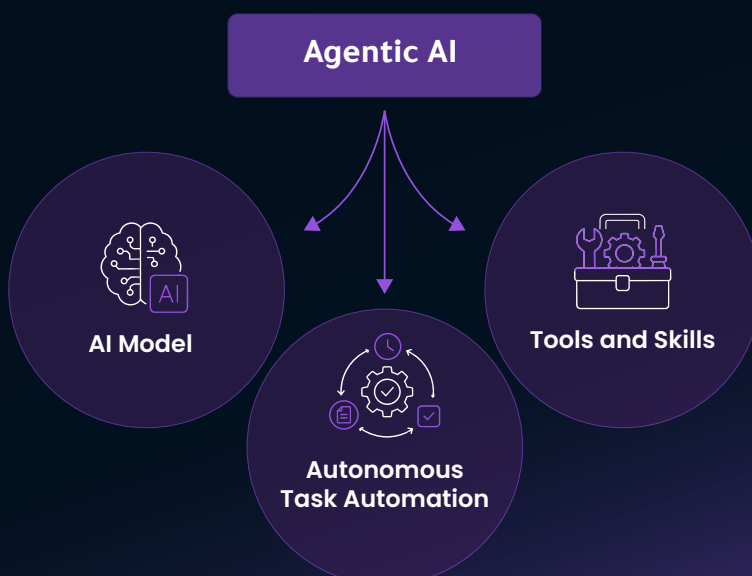


Self-Learning:

Installing new software on the system that requires specific skills to operate.

The Agentic AI trains and teaches itself how to use these programs based on the permissions granted to it.

Components of Agentic AI and How It Works



We can understand how it works as follows:

- ✓ **Goal Understanding & Planning:** Understanding requirements, analyzing context, and breaking tasks into smaller steps using the AI Model.
- ✓ **Tools & Skills Utilization:** Employing necessary tools and skills to complete the task, such as utilizing email software (e.g., Microsoft Outlook).
- ✓ **Task Automation:** Periodically checking email and responding (via formal letters or emails) according to instructions using skills, task automation, and continuous self-learning.

In Simple Terms:



Traditional AI (Chatbots)

provides you with answers.



Agentic AI

can execute actions.

Technical Threats in Agentic AI Components

When Agentic AI is granted full permissions for autonomous action, the potential attack surface expands significantly across its components:



AI Model:

AI models suffer from producing inaccurate or incorrect outputs presented with high confidence (hallucinations). In agentic systems, these hallucinations and incorrect details translate directly into real, automated actions and commands.



Tools and Skills:

Granting an agent broad system permissions and execution rights means that any compromise of the agent or a successful Prompt Injection attack could allow an attacker to exploit these permissions to access systems and execute unauthorized commands, expanding the scope and impact of the attack.



Autonomous Task Automation:

Agents operate at high speeds. In the absence of human oversight, the system might rapidly repeat an error or malicious command, potentially resulting in data deletion, service disruption, or inflated operational costs.

How Attackers Exploit Agentic AI

Attackers may abuse the permissions granted to Agentic AI or manipulate the content it processes to trick it into executing harmful actions or accessing sensitive data without the user's knowledge through:



Running Untrusted Skills:

Some agents rely on third-party skills or plugins to extend their capabilities. Running a skill from an untrusted source is equivalent to executing malicious software on a work device.



Indirect Prompt Injection:

The agent reads external data like emails and websites. Attackers can hide malicious instructions within this data, causing the agent to treat them as legitimate commands and execute them automatically.



Data Leakage:

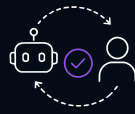
The system may require login credentials or API keys to perform tasks. If not properly secured, these credentials can be exposed or stolen.

Use Agentic AI with Caution

While Agentic AI can accomplish tasks rapidly, its safe deployment requires continuous verification and strict security controls:



Grant the least privilege necessary to accomplish assigned tasks and run the agent in a secure environment.



Enforce Human-in-the-Loop manual approvals prior to executing sensitive operations.



Verify the trustworthiness of AI models and skill sources before adopting or integrating them.



Monitor system execution logs continuously and report any unexpected behavior.



Cyber Knowledge:

Do you know what is **Local AI / LLM**?

Local LLMs are models that run locally on private devices and servers without requiring an internet connection during use. While this helps protect privacy, it does not make the system totally secure. Local deployments introduce the following challenges:



Downloading open-source models from untrusted sources.



Utilizing uncensored LLMs that do not enforce safety controls.



Difficulty in monitoring usage if deployed outside official IT channels (Shadow AI).



Requires High-end hardware and secure infrastructure.

