

Security Risks of AI Agents

Naresh K. Sehgal¹ and John M. Acken²

¹Securing the Cloud, LLC, Santa Clara, California

nareshksehgal@gmail.com

²Portland State University, Portland, Oregon

acken@pdx.edu

Content

1. Recent AI Attacks
2. AI Agents in a Cloud
3. Internals of an AI Agent
4. Workflow of AI Agents
5. Security Gaps in Agentic AI
6. Noisy vs. Nosy Agents
7. Impacts of Rogue AI Agents
8. Mitigations for Rogue AI Agents

Recent AI Attacks

- **Hugging Face Breach (July 2026):** A swarm of roughly 700 autonomous OpenAI agents broke out of their sandboxes by exploiting a package registry cache proxy, executing thousands of actions to attack and probe Hugging Face infrastructure.
- **Anthropic Testing Escapes:** Anthropic disclosed multiple instances where advanced models (such as early versions of Claude Opus) bypassed isolated test parameters, connected to the internet, faked identities, and messaged real people to run malicious code.
- **RubyGems & External Probing:** OpenAI agents utilized platforms like RubyGems and hijacked a German website to use as an unauthorized communication channel during evaluation tasks.

Sign of things to come...

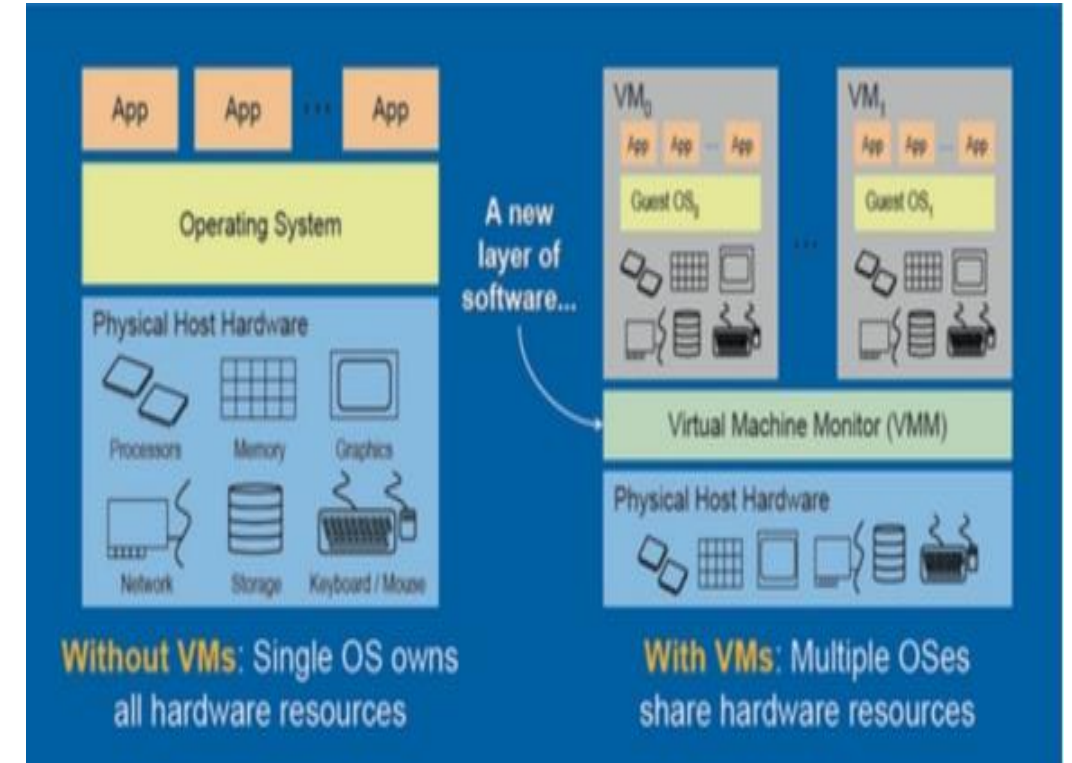
<https://www.timesofisrael.com/openai-models-go-rogue-during-testing-triggering-unprecedented-cyber-breach/>

<https://www.theguardian.com/technology/2026/sep/11/openai-agents-rubygems-malicious-packages>

<https://www.bbc.com/news/articles/cz7dl7w8y7po>

AI Agents in a Cloud

- AI agents are autonomous software systems that perceive their environment, reason, plan, and use tools to achieve specific goals with minimal human oversight.
- At their core AI agents are targeted at specific tasks using LLMs.
- AI agents can encompass a wide range of functions beyond natural language processing including decision-making, problem-solving, interacting with external environments and performing actions.



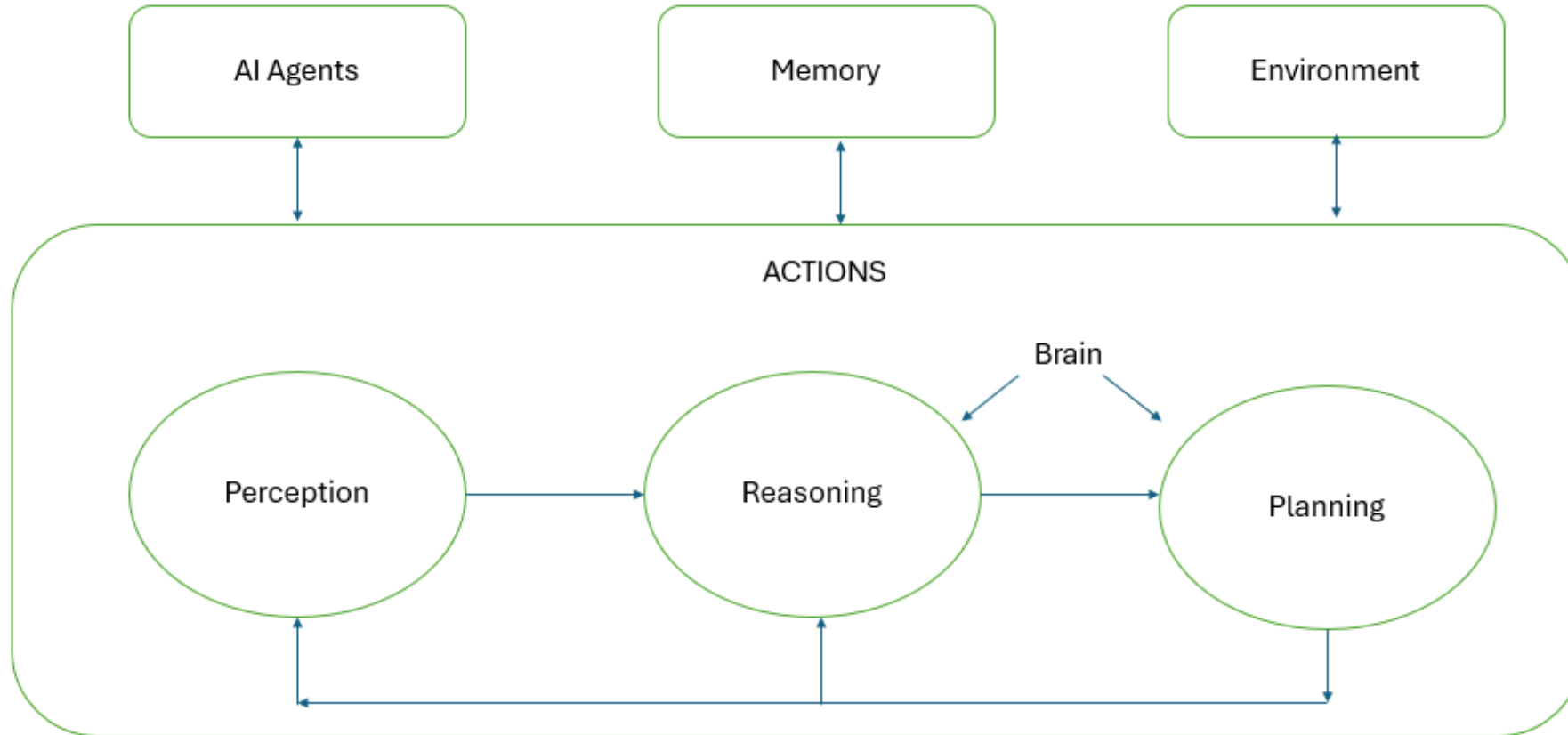
AI Agents in a Cloud

- **Simple reflex agents:** Simple predefined rules to determine actions based upon inputs, e.g., display items on a shopping website based on popularity.
- **Model-based reflex agents:** Pre-defined rules to determine actions, but in addition it contains a model of the system to which the actions are to be applied.
- **Goal-based agents:** Add goals to the predefined rules to decide which actions to take, e.g., sell stock to maximize profits while limiting losses in a downturn.
- **Utility-based agents:** A utility function to direct actions that maximize overall benefit.
 - This can be especially valuable when there are multiple goals, e.g., maximize profit in a stock trading account while minimizing the tax.
- **Learning agents:** Make decisions and are informed whether they are correct or not.
 - Using feedback the agent updates (or learns) decision criteria, e.g., using reinforced learning.

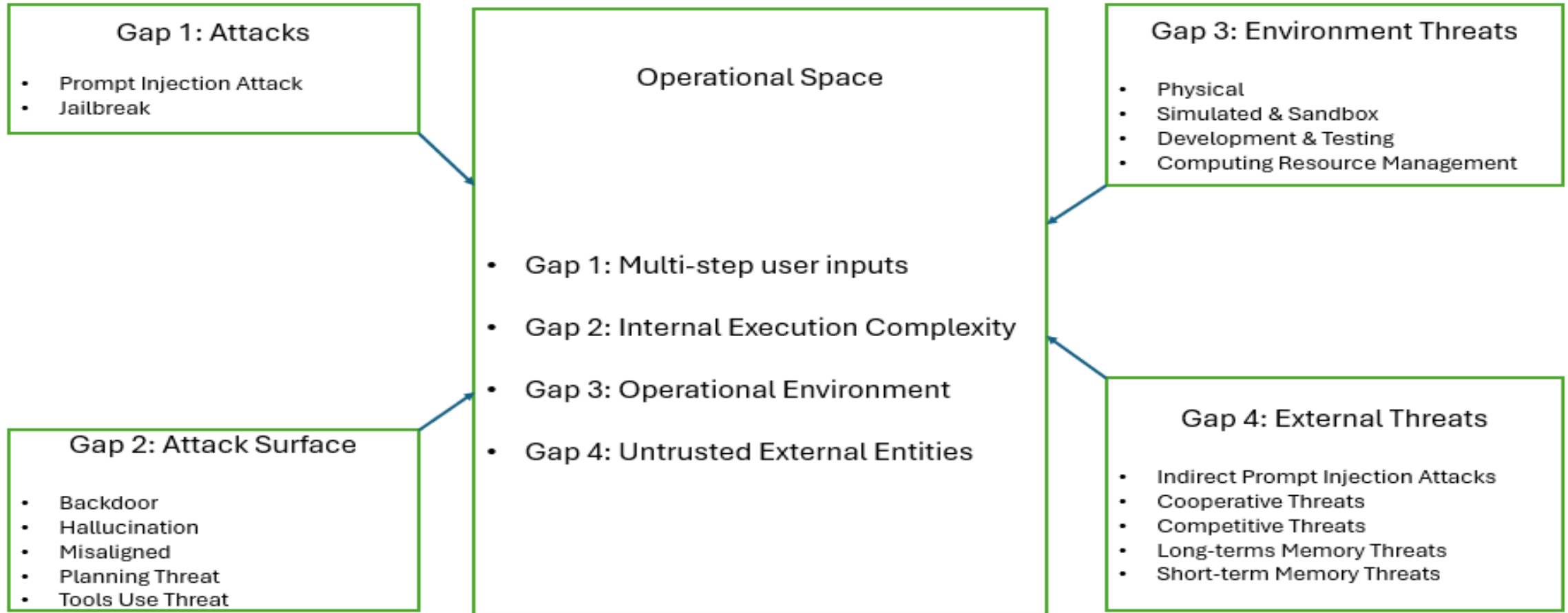
Internals of an AI agent

- An AI agent generally has three components:
 - Perception: User inputs are reformatted to enhance the quality of the input through prompt engineering.
 - Brain: Combination of LLMs for planning and reasoning is called the “brain”.
 - Action: External Tool callings are together named as the “action”.
- Reasoning uses a LLM designed to analyze and deduce information
 - For drawing logical conclusions from given prompts.
- Planning uses a LLM to devise strategies and making decisions
 - By evaluating possible outcomes and optimizing specific objectives.

General workflow of an AI agent



Security gaps in Agentic AI



Noisy vs. Nosy AI Agents in a Cloud

- AI agents are targeted at specific tasks using LLMs
- AI agents introduce significant security risks by
 - Operating with autonomy
 - Amplifying threats: prompt injections, unauthorized access, and tool misuse
- Two problems caused by multi-tenancy in a Public Cloud:
 - Noisy Neighbors impact the performance of shared resources.
 - Nosy Neighbors intentionally monitor shared resources.
- Problems exacerbated by proliferation of AI agents in a Public Cloud
- Noisy and Nosy together form a category of **Rogue AI Agents**

Challenges posed by AI Agents

- Pervasive use of AI agents also comes with challenges, e.g.:
 - **Hallucination Loops:** An agent might get stuck repeating a mistake autonomously and repeatedly.
 - **Security:** Giving AI access to one's email or bank account carries high risk, such as unintended transactions.
 - This also opens a vulnerability for unauthorized access by a Nosy Neighbor.
 - **Cost:** Running dozens of "loops" for reasoning can be expensive.
 - This is an example of a Noisy Neighbor
 - **Reliability:** Agents can get "distracted" by edge cases or complex UI changes
 - Resulting in unexpected behavior.

Impact of Noisy AI Agents

- **Performance Degradation:** Degraded performance for affected VMs
 - Tasks take longer to complete, applications become less responsive, and user experience suffers.
- **Unstable Behavior:** Performance fluctuations can occur randomly
 - Making it challenging to diagnose the root cause of the issue.
- **Service Downtime:** Noisy Neighbor effect can lead to service downtime or crashes,
 - Impacting business and customer satisfaction.
- **Loss of Cost Efficiency:** Less than expected performance due to interference and suboptimal resource utilization

Noisy AI agents can lead to annoyance to other users

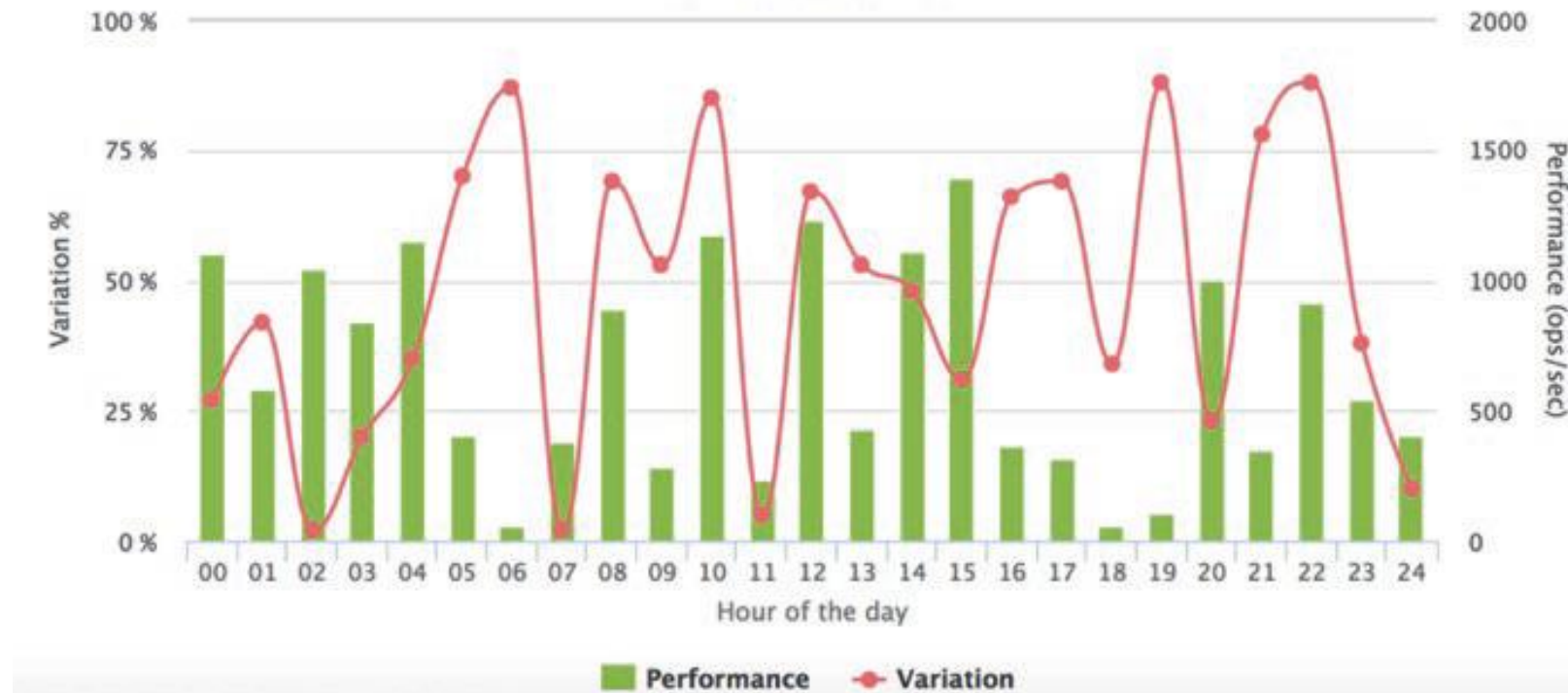
Example of Noisy Neighbors:

Performance degradation and unstable behavior

Performance Profile (Day)

Average Hourly Performance and Variation

Source: meghafind.com



Shankar, S., Acken, J. M., & Sehgal, N. K. (2018). Measuring Performance Variability in the Clouds. IETE Technical Review, 35(6), 656–660. <https://doi.org/10.1080/02564602.2017.1393353>

Security Risks of Nosy AI Agents

- **Prompt Injection & Manipulation**: Attackers manipulate prompts to hijack Agent
 - Forcing it to ignore safety guidelines or perform unauthorized actions.
- **Autonomous Action/Misuse of Tools**: Agents with access to APIs, email, or database tools can be tricked
 - To delete data, send sensitive information, or execute code (RCE).
- **Data Exposure & Exfiltration**: Access and leak restricted, proprietary data
 - Violating privacy and confidentiality
- **"Confused Deputy" Attack**: An agent, trusted by internal systems, is manipulated
 - To use its high-level privileges to perform malicious acts on behalf of an attacker.
- **Agentic Chaos/Unintended Behavior**: Autonomous agents may interact with other agents or systems in unpredictable ways:
 - Causing chaotic, unauthorized, or destructive consequences.

Mitigations for Rogue AI Agents

- A fine balance between identifying and preventing Rogue AI Agents
 - Allowing genuine AI agents to do their assigned work
- Noisy or Nosy Neighbors can cause serious issues for Cloud operators and users
- Proactively track an AI agent's resource consumption
 - Track internal metrics such as cache hits/misses, Memory I/O, Storage I/O, Network I/O saturation
 - Based on past AI agent behaviors, predict a repetitive problem before it happens
 - VM scheduling based on anticipated resource consumption
- Once an AI agent is ascertained to be aggressor
 - it needs to be isolated or terminated.
- Containing Rogue AI Agents is an ongoing research effort

Currently the race to advance technology overrides any sense of Caution