Agentic AI: Cybersecurity Analyst or Cyber Threat?

Shapla Khanam
Dept. AI FCSIT, UM

Agentic AI-Driven Cybersecurity

The next cyber threat isn't a human hacker anymore, but an autonomous AI agent can plan, reason, and attack at machine speed. With the rapid development of AI technologies, cyber-attacks are becoming more sophisticated and increasingly AI-driven. Conventional cybersecurity tools struggle to keep pace. Millions of cyber alerts overwhelm the Security Operations Centers (SOCs) every hour. The global shortage of cybersecurity analysts continues to grow, making it increasingly difficult to manage these alerts.

Moreover, traditional AI-based cybersecurity models function as black boxes, making their detection and response decisions difficult for security analysts to understand and trust. This has moved the attention to Agentic AI, the next generation of AI.

Unlike chatbots or GenAI that simply reply to prompts, Agentic AI offers new opportunities for autonomous cybersecurity. Agentic AI enables autonomous AI agents to observe network events, reasons about attacks, make autonomous decisions and coordinate response. Example, an AI agent is able to investigate malicious logins, correlate security alerts, consult threat intelligence, recommend appropriate & timely mitigation plans, and even execute defensive actions in seconds.

AI Autonomy Risks

Human–AI Collaboration

Agentic AI is significantly changing how a SOC operate. Agentic AI is capable of enhancing threat detection, reducing alert exhaustion, accelerating response times, and constant monitoring. Leading cybersecurity organizations are already incorporating AI agents into their tech platforms, while projects such as the DARPA AI Cyber Challenge (AIxCC) are advancing autonomous vulnerability analysis and remediations.

Nevertheless, this autonomy also establishes new dangers. While Agentic AI is capable of defending systems, it can also become smart targets, or perhaps mighty tools for hackers. Cybersecurity analysts and researchers have discovered various evolving threats like agent hijacking, prompt injection, memory and tool poisoning among others.

Malicious agents take control over or compromise AI agents to execute unintended actions. These emerging threats have driven organizations such as NIST, OWASP, MITRE, ENISA, and CISA to foster direction for the safe and trustworthy deployment of AI agents.

Contact at: shapla.k@um.edu.my