MIT Professional Education

Applied Agentic AI for Organizational Transformation

Glossary

Module 4

Confidence Score (AI): A metric provided by some AI systems that indicates the likelihood that the model's output is correct. This score can help assess reliability but is not always available.

Context Window (LLM): The amount of information a language model can retain and consider when generating a response. A limited context window can result in the omission of relevant details in complex tasks.

Cyber Kill Chain: A cybersecurity framework that divides an attack into seven stages, from reconnaissance to data exfiltration, and provides opportunities for detection and response at each phase.

Deepfake: AI-generated audio or video content that convincingly mimics real individuals. Deepfakes support impersonation, disinformation, or manipulation and are often difficult to detect without specialized tools.

Exfiltration: The final phase in many cyberattacks in which the attacker extracts sensitive data from the compromised system, often resulting in financial or reputational harm.

Impersonation Attack: A cybersecurity tactic in which attackers use AI tools such as chatbots, voice clones, or fake personas to pose as trusted individuals or brands in order to manipulate targets or steal data.

Least-Privilege Access: A security principle that ensures AI agents or users are granted only the minimum permission required to perform their tasks, reducing the risk of misuse or compromise.

Phased Deployment (AI): A strategy for gradually implementing AI systems, starting with low-risk use cases and increasing autonomy over time. This method helps reduce risk and improve oversight.

Prompt Injection: An exploit in which attackers insert hidden or malicious instructions into prompts directed at AI systems, causing them to act in unintended or harmful ways.

Reconnaissance (Cybersecurity): The initial stage of a cyberattack in which the attacker gathers information about the target organization, often using public sources, to identify vulnerabilities.

Remote Access Trojan (RAT): Malware that enables an attacker to gain unauthorized control over a computer system, often without the user's knowledge.

Sandboxing: A cybersecurity technique used to isolate systems or processes, including AI agents, to prevent access to critical resources or limit harm if compromised.

Spoofing: The act of disguising communication from an unknown or untrusted source as being from a known, trusted source. AI can increase the scale and believability of spoofing attempts.

Synthetic Disinformation: False or misleading information created and distributed using AI-generated content, such as deepfake videos or voice clones, with the intent to deceive or manipulate.

Weaponization (Cybersecurity): A stage in the Cyber Kill Chain in which attackers create malicious payloads, such as infected documents or prompt-based exploits, to target a specific individual or system.