Infopercept | INSENSE

Infopercept 2026 Threat Predictions:

Attacks on AI & Attacks Using AI

Infopercept initiated this 2026 threat research because, for the first time in cybersecurity history, attackers and defenders are powered by the same engine — artificial intelligence. The rise of GenAI has erased traditional asymmetries of capability: adversaries no longer need elite skills to launch sophisticated attacks, and defenders can no longer rely on rarity of expertise as a barrier. This shared access to the same AI powerhouse has fundamentally changed the dynamics, velocity, and volume of cyberattacks.

The research is divided into two major portions

Attacks on AI and Attacks with AI

reflecting how artificial intelligence has simultaneously become both the target and the weapon in the modern cyber battlefield. Infopercept's predictions aim to help organizations anticipate this new landscape where the number of lethal adversaries will multiply, and the line between automation and autonomy in both offense and defense will blur faster than ever before.

A. Attacks on AI

1. GenAI democratization → surge in data poisoning & software supply chain compromise

Prediction Summary:

By 2026, GenAI-assisted coding will allow anyone — even non-developers — to produce functional applications. This low barrier will multiply insecure code and data pipelines. Threat actors will exploit this by injecting poisoned datasets, seeded insecure prompts, and malicious code templates into shared repositories, enabling large-scale AI-assisted supply chain compromise.

Possible Attacks:

- Poisoned datasets uploaded to internal or public repositories that subtly introduce exploitable logic or backdoors into GenAI-generated applications.
- Prompt-injection templates that embed obfuscated malicious payloads in generated code or configuration files.
- Trojanized plug-ins for AI code assistants that insert exfiltration routines or dependency confusion vectors.

Impact:

- Compromise of entire app stacks developed through AI assistants.
- Broader exploitation surface in low-code / no-code workflows.

Detection Indicators:

- Common vulnerable code patterns emerging across independent projects.
- Abnormal API calls in newly created applications.

Mitigations:

- Signed and versioned prompt/data repositories.
- Mandatory static analysis and software composition analysis on all AI-generated code.
- Training data validation pipelines and provenance tracking.

2. Model Context Protocol (MCP) expansion → new lateral and contextual manipulation attacks

Prediction Summary:

The growing adoption of Model Context Protocol (MCP) — used to connect LLMs with real-time context (files, tools, APIs) — introduces a new attack plane: manipulation of the context layer itself. Since MCP acts as the bridge between models and operational environments, compromising it means controlling what the AI "sees" and "believes."

Possible Attacks:

- **Context poisoning:** Attackers inject malicious or misleading information into context sources (e.g., documents, APIs, or databases) that MCP fetches.
- **Context hijacking:** Adversaries modify MCP configurations or endpoints to redirect models to attacker-controlled contexts.
- **Permission overreach:** Exploiting poorly scoped MCP connectors to access sensitive internal systems through the model's context window.
- **Reflection attacks:** Recursive loops triggered between interconnected MCPs (AI-to-AI loops) leading to data leakage or system overload.

Impact:

- Large-scale misinformation or model misbehavior.
- Compromised decision-making pipelines in enterprises using MCP for autonomous operations.

Indicators:

- Models producing contradictory or manipulated outputs traced to altered context files.

- Sudden changes in MCP connector behavior (unexpected file/API access).

Mitigations:

- Cryptographic signing of MCP contexts and connector manifests.
- Zero-trust access enforcement on all MCP connectors.
- Continuous validation of retrieved context content (e.g., checksum and schema verification).
- Sandbox and test environments for MCP updates before production rollout.

3. Multi-LLM usage → gateway bypass and adversarial routing

Prediction Summary:

Organizations will increasingly deploy multiple LLMs behind LLM gateways for cost, specialization, or compliance reasons. Adversaries will mimic how they once bypassed firewalls — crafting indirect prompts, covert payloads, and rogue connectors to evade LLM gateways and exfiltrate or poison information.

Possible Attacks:

- **Gateway evasion:** Crafting adversarial prompts that split malicious tasks across different LLMs to bypass policy enforcement.
- **Shadow routing:** Using unsanctioned LLMs that connect through hidden connectors outside the gateway.
- **Cross-model leakage:** Prompting one model to reveal sensitive outputs generated by another model through chain-of-thought inference.

Impact:

- Massive data leakage and compliance breaches.
- Loss of visibility over LLM interactions.

Indicators:

- Anomalous token bursts or inter-LLM data transfers.
- API calls to non-approved LLM endpoints.

Mitigations:

- Strict allow-list enforcement for LLM endpoints and connectors.
- Policy-based output sanitization and contextual watermarking.
- Runtime inspection and anomaly detection for LLM cross-calls.

4. Proliferation of SOC agents → agent poisoning and orchestration hijack

Prediction Summary:

By 2026, Security Operations Centers (SOCs) will rely on dozens of autonomous security agents for alert triage, incident response, and threat hunting. These agents themselves will become high-value targets. Attackers will poison their data sources, exploit API keys, or inject malicious playbooks to manipulate SOC decisions.

Possible Attacks:

- **Playbook poisoning:** Adversaries modify automation workflows so that agents disable sensors or ignore certain alerts.
- **Telemetry manipulation:** Injecting fake events into agents' data sources to trigger false quarantines or hide real intrusions.
- **Agent hijacking:** Compromise of orchestration APIs allowing attackers to push new malicious instructions to all agents.

Impact:

- Automated destruction of evidence or mass quarantines of clean systems.
- Loss of trust in AI-driven response.

Indicators:

- Playbook changes outside change windows.
- Agents performing unauthorized network modifications or file deletions.

Mitigations:

- Digitally sign and audit all playbooks.
- Human approval for destructive automated actions.
- Immutable logging and real-time playbook integrity checks.

5. Identity layer with AI agents → token forgery and privilege chaining

Prediction Summary:

As IAM layers integrate AI agents for decision-making (auto-provisioning, risk scoring), attackers will exploit delegation tokens, impersonation routes, and identity-based agents to escalate privileges. The identity fabric becomes a lattice of agents — and each one is an attack entry.

Possible Attacks:

- **Token replay or theft:** Extracting agent credentials from caches or memory.
- **Agent impersonation:** Registering fake AI identity agents mimicking legitimate automation accounts.
- **Privilege chaining:** Using one agent's delegated authority to pivot laterally across services.

Impact:

- Stealthy privilege escalation and long-term persistence.

Indicators:

- Unusual token usage from new or low-activity agents.
- Identity graph anomalies — agents calling APIs unrelated to their assigned roles.

Mitigations:

- Short-lived credentials with enforced rotation.
- Behavioral baselines for agent identities.
- Continuous attestation of active AI agents.

6. AI-based security testing poisoned → weakened SDLC

Prediction Summary:

If AI-based code testing, fuzzing, or vulnerability scanners are poisoned or manipulated, they may overlook or misreport flaws — creating a false sense of security. Attackers can bias models to ignore specific vulnerabilities or insert insecure "auto-fixes."

Possible Attacks:

- Data poisoning in AI testing datasets (teaching the AI that vulnerable patterns are safe).
- Exploit suggestion bias — AI generates "secure" configurations that disable security features.
- Supply-chain compromise of AI testing APIs or model weights.

Impact:

- Systemic production of vulnerable code.
- Recurrent exploitation of the same flaw types.

Mitigations:

- Human oversight and cross-validation of AI test results.
- Use of benchmark vulnerability corpora for continuous model evaluation.
- Isolation and signature verification of testing model updates

7. MITRE adapting its framework for AI → shift in attack mapping

Prediction Summary:

The expansion of MITRE's ATT&CK framework to cover AI-specific tactics (prompt injection, model evasion, data poisoning, model exfiltration) will professionalize both defense and offense.

Adversaries will use these standardized patterns to refine and automate their own playbooks, similar to how they exploited MITRE ATT&CK initially.

Prediction:

- Emergence of "AI-aware" malware built specifically to avoid AI monitoring controls.
- Red-teaming tools that emulate AI attacks using the MITRE AI matrix as a blueprint.

Mitigations:

- Align detection logic with new MITRE AI techniques.
- Continuous update of SOC content to track AI attack evolution.

8. On-prem / Air-gapped AI adoption → erosion of isolation

Prediction Summary:

Traditional "secure" environments (on-prem, air-gapped critical infrastructure) will integrate AI for predictive maintenance and anomaly detection. This requires data import/export bridges, breaking historical isolation. These bridges become prime infiltration channels.

Possible Attacks:

- Infected datasets or models imported via USB or controlled synchronization channels.
- Model update packages containing malware that executes during deployment.
- Insider manipulation of transfer workflows.

Impact:

- Breach of previously untouchable critical systems.
- Potential safety implications in industrial or healthcare OT.

Mitigations:

- One-way transfer controls (data diodes) with cryptographic verification.
- Mandatory digital signatures and attestation for imported models.
- Separation of AI computation nodes from core OT control systems.

9. Shadow AI instances → invisible backdoors

Prediction Summary:

Unapproved AI deployments — "Shadow AI" — will proliferate as employees or departments spin up private LLMs and agents. These rogue instances will bypass governance and become exfiltration and poisoning vectors.

Possible Attacks:

- Shadow AI tools sending sensitive data to third-party APIs.
- Poisoned or malicious AI apps spreading false outputs internally.

Impact:

- Data leakage, compliance failures, and unmonitored threat vectors.

Mitigations:

- Continuous discovery of unauthorized AI endpoints.
- DLP and firewall rules blocking unapproved AI traffic.
- Policy enforcement via LLM gateway and identity-based controls.

10. Agentic malware/ransomware → autonomous threat evolution

Prediction Summary:

By 2026, AI-powered malware will evolve into agentic malware — autonomous software capable of goal-driven behavior, learning from failed attempts, and making independent operational decisions.

Possible Attacks:

- Agentic ransomware identifying critical assets autonomously before encryption.
- Autonomous extortion bots that communicate, negotiate, and escalate without C2 servers.
- Self-propagating agents that "collaborate" to sustain persistence.

Impact:

- Ultra-fast, adaptive attacks that outpace human response.

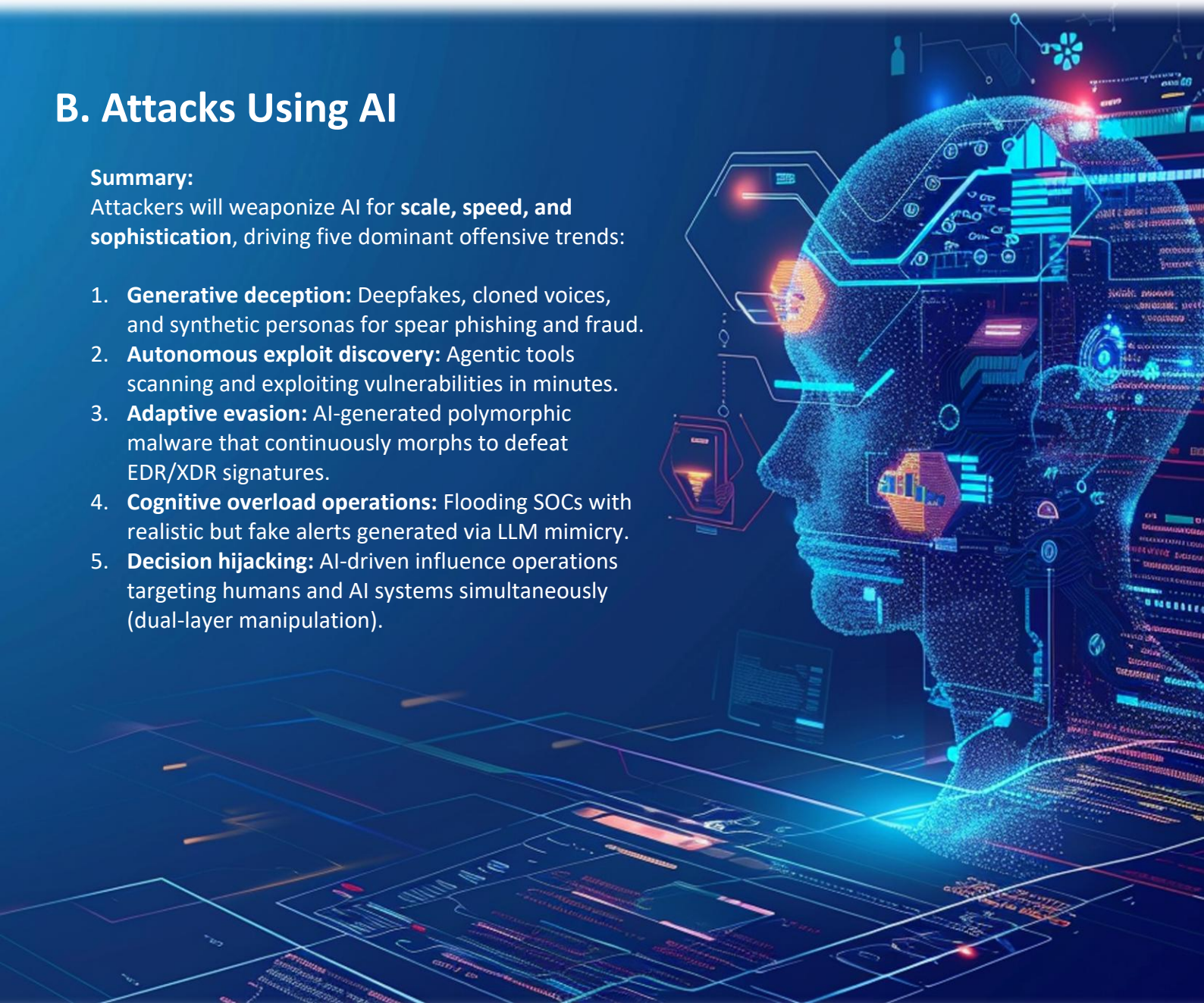
Mitigations:

- AI behavior anomaly detection (learning deviation analysis).
- Micro-segmentation and deception infrastructure (decoy assets).
- Immutable backups and continuous behavioral isolation

B. Attacks Using AI

Summary:
Attackers will weaponize AI for **scale, speed, and sophistication**, driving five dominant offensive trends:

1. **Generative deception:** Deepfakes, cloned voices, and synthetic personas for spear phishing and fraud.
2. **Autonomous exploit discovery:** Agentic tools scanning and exploiting vulnerabilities in minutes.
3. **Adaptive evasion:** AI-generated polymorphic malware that continuously morphs to defeat EDR/XDR signatures.
4. **Cognitive overload operations:** Flooding SOCs with realistic but fake alerts generated via LLM mimicry.
5. **Decision hijacking:** AI-driven influence operations targeting humans and AI systems simultaneously (dual-layer manipulation).



Key Takeaways for 2026

	Dominant Threats	Drivers	Required Defenses
Attacks on AI	Context poisoning, model theft, agent hijacking, AI supply-chain compromise	MCP, Multi-LLM ecosystems, shadow AI	Signed context, zero-trust connectors, gateway inspection, model attestation
Attacks using AI	Agentic malware, GenAI-powered phishing, autonomous exploitation	Adversarial AI evolution, low-code proliferation	AI-driven threat detection, deception tech, multi-layer XDR correlation
Cross-impact	Erosion of trust in AI automation and security analytics	Increased reliance on AI in SOCs and SDLC	Human-AI hybrid governance and validation loops

About Infopercept's Threat Research Team

Infopercept's Threat Research Team brings together offensive, defensive, and AI security specialists who continuously study emerging attack behaviors across the cyber kill chain. The team operates at the intersection of red teaming, threat intelligence, and platform engineering — combining real-world adversarial simulation with data-driven defense insights from the Invinsense platform. Their research focuses on how evolving technologies such as GenAI, autonomous agents, and adaptive malware are transforming both attack and defense surfaces. Each year, the team publishes forward-looking predictions to help organizations prepare for the next wave of cyber and AI-driven threats.

Disclaimer & Copyright Notice

This publication represents Infopercept's independent threat research and forward-looking analysis based on publicly available data, internal threat intelligence, and expert interpretation. The scenarios and predictions described are intended to inform cybersecurity awareness and strategic planning. They do not rely on or reproduce any proprietary data or content from third parties.

About Infopercept - Infopercept is one of the fastest-growing comprehensive cybersecurity companies in India, serving global clients. It provides platform led managed security services that covers all areas of cybersecurity, including defensive, offensive, detection and response, and security compliance. Infopercept has its own cybersecurity platform, 'Invinsense,' which integrates tools such as SIEM, SOAR, EDR, deception, offensive security, and compliance tools. Its cybersecurity and MDR services include dedicated teams of experts, ensuring that organizations have 24x7 cybersecurity operations support.

Imprint

Infopercept Consulting Pvt. Ltd.

Publisher Address

3rd floor, Optionz Complex, CG Rd, Opp. Regenta Hotel, Navrangpura, Ahmedabad, Gujarat 380009, INDIA

Contact

sos@infopercept.com
www.infopercept.com