DropTicks AI Article

AI Security for Agents: Treat Model Outputs as Untrusted Input

AI Security | Mar 13, 2026

A security checklist for tool-using agents, MCP servers, prompt injection, secrets, file access, and approval boundaries.

Article

Agentic AI increases the value of security basics. The model can read user input, retrieve documents, call tools, write files, and trigger workflows. That means model output should be treated as untrusted input.

A secure agent design includes:

1. Least-privilege tools.
2. Explicit allowlists for actions.
3. Human approval for high-impact changes.
4. No secrets in prompts or retrieved context.
5. Validation on tool arguments before execution.
6. Logging for every tool call.
7. Isolation for code execution.
8. Separate read and write permissions.
9. Tests for prompt injection and tool misuse.

Security failures often happen at the connector layer: filesystem access, database queries, shell execution, or unchecked user-controlled metadata. The model should propose an action; the application should validate whether that action is allowed.

For DropTicks systems, security is not a late-stage checklist. It is part of the system design: define who can do what, which data can be read, which actions need approval, and how every important action is audited.

Web link

https://www.dropticks.com/articles/ai-security-for-agents-treat-model-outputs-as-untrusted-input

Related links

https://modelcontextprotocol.io/docs/getting-started/intro
https://openai.github.io/openai-agents-python/
https://docs.litellm.ai/docs/