How does FortressAI know an AI tool hasn't been tampered with?

Short answer: Before granting any AI tool access to data, FortressAI assesses whether that tool has been tampered with — checking the integrity of its process and loaded libraries through the same memory behavioral analytics the rest of the platform runs on. An approved tool that has been compromised does not get your data simply because it's on the allowed list.

Approval is not the same as trust

Most AI governance stops at the allow-list: is this application permitted? That's necessary but not sufficient, because an application's identity and its *integrity* are different questions.

`claude.exe`, a Copilot client, a ChatGPT desktop app, or a browser running Gemini can all be legitimately approved — and all be exploited. An attacker who compromises an approved AI client inherits whatever data access that client was granted. An allow-list alone hands it over.

FortressAI therefore evaluates two things before data access:

1. **Is this tool authorized?** (per-tool assessment and policy)
2. **Is this tool still itself?** (has the process or its libraries been tampered with)

Both must hold.

What "tampered with" means here

The check draws on memory analytics — the same foundational sensor layer behind data and identity protection. It looks at the state of the running program and its loaded libraries for signs of injection, in-memory modification, or exploitation.

This matters because modern tradecraft specifically targets trusted processes. Code injected into an approved application inherits its permissions and its reputation. Checking the file on disk proves nothing about what the process has become in memory.

What happens on a failed check

The response follows the same graduated model used across the platform:

- **Authorized and untampered, within policy** — access proceeds.
- **Authorized and untampered, but exceeding policy** — blocked, redacted, or given fake data, depending on your rules. The tool keeps running.
- **Tampered or unknown** — rejected or suspended, per the behavioral engine and your settings.

The practical effect: a compromised AI client is treated as an attacker, not as an approved application having a bad day.

Revision #3
Created 2026-07-20 19:24:00 UTC by Dennis Underwood
Updated 2026-07-21 19:32:15 UTC by Dennis Underwood