Woher weiß FortressAI, dass ein KI-Tool nicht manipuliert wurde?

Kurze Antwort: Bevor einem KI-Tool Zugriff auf Daten gewährt wird, prüft FortressAI, ob dieses Tool manipuliert wurde – die Integrität seines Prozesses und der geladenen Bibliotheken wird über dieselbe verhaltensbasierte Speicheranalyse geprüft, auf der auch der Rest der Plattform basiert. Ein zugelassenes Tool, das kompromittiert wurde, erhält Ihre Daten nicht einfach nur deshalb, weil es auf der Zulassungsliste steht.

Genehmigung ist nicht dasselbe wie Vertrauen

Die meisten Ansätze zur KI-Governance enden bei der Zulassungsliste: Ist diese Anwendung erlaubt? Das ist notwendig, aber nicht ausreichend, denn die Identität einer Anwendung und ihre *Integrität* sind unterschiedliche Fragen.

`claude.exe`, ein Copilot-Client, eine ChatGPT-Desktop-App oder ein Browser mit Gemini können alle rechtmäßig genehmigt sein – und alle ausgenutzt werden. Ein Angreifer, der einen genehmigten KI-Client kompromittiert, erbt jeden Datenzugriff, der diesem Client gewährt wurde. Eine Zulassungsliste allein übergibt diesen Zugriff einfach.

FortressAI prüft daher zwei Dinge, bevor Datenzugriff gewährt wird:

1. **Ist dieses Tool autorisiert?** (Bewertung und Richtlinie pro Tool)
2. **Ist dieses Tool noch es selbst?** (wurde der Prozess oder seine Bibliotheken manipuliert)

Beides muss zutreffen.

Was „manipuliert" hier bedeutet

Die Prüfung stützt sich auf Speicheranalysen – dieselbe grundlegende Sensorebene, die auch dem Daten- und Identitätsschutz zugrunde liegt. Sie untersucht den Zustand des laufenden Programms und seiner geladenen Bibliotheken auf Anzeichen von Injektion, Speichermodifikation oder Ausnutzung.

Dies ist wichtig, weil moderne Angriffstechniken gezielt vertrauenswürdige Prozesse ins Visier nehmen. Code, der in eine genehmigte Anwendung eingeschleust wird, erbt deren Berechtigungen und deren Reputation. Die Überprüfung der Datei auf der Festplatte belegt nichts darüber, wozu der Prozess im Arbeitsspeicher geworden ist.

Was bei einer fehlgeschlagenen Prüfung passiert

Die Reaktion folgt demselben abgestuften Modell, das in der gesamten Plattform verwendet wird:

- **Autorisiert und unmanipuliert, im Rahmen der Richtlinie** — Zugriff wird gewährt.
- **Autorisiert und unmanipuliert, aber über die Richtlinie hinausgehend** — je nach Ihren Regeln blockiert, geschwärzt oder mit gefälschten Daten versehen. Das Tool läuft weiter.
- **Manipuliert oder unbekannt** — abgelehnt oder ausgesetzt, gemäß der Verhaltensengine und Ihren Einstellungen.

Die praktische Auswirkung: Ein kompromittierter KI-Client wird wie ein Angreifer behandelt, nicht wie eine genehmigte Anwendung, die gerade einen schlechten Tag hat.

Revision #1
Created 2026-07-24 04:36:46 UTC by Dennis Underwood
Updated 2026-07-24 04:36:46 UTC by Dennis Underwood