🗓 12.02.2026 — Gray Zone – Was Anthropic wirklich berichtet

Posted on Februar 12, 2026 by Redaktion-AnalyseTeam

Anthropic stuft sein Modell Claude Opus 4.6 in eine „Gray Zone“ ein: In Tests zeigte es eine erhöhte Anfälligkeit für Missbrauch, unter anderem bei der Simulation chemischer Bedrohungsszenarien. Das Unternehmen bewertet das Sabotagerisiko als „sehr gering, aber nicht vernachlässigbar“ und reagiert mit zusätzlichen Sicherheitsauflagen im Rahmen seiner Responsible Scaling Policy.

Die eigentliche Frage lautet jedoch nicht, ob Risiken existieren – sondern wie transparent mit ihnen umgegangen wird. Entscheidend bleibt die **Verantwortungsrückbindung**: klare Zuständigkeiten, überprüfbare Schutzmechanismen und die Bereitschaft, Entwicklung notfalls zu bremsen.

🔗 Gray Zone – Was Anthropic wirklich berichtet (und was daraus gemacht wird)