SÉCURITÉ & IA

Décoller sans savoir atterrir

Le Komet, avion-fusée de 1944, n'avait pas de train d'atterrissage. En 2026, des modèles d'IA sortis de leur bac à sable posent la même question : ce qui manque, c'est tout ce qui permet de revenir au sol.

Publié le 26 novembre 2026 — Imprimé le 27 novembre 2026



1944 hatte das schnellste Flugzeug am Himmel kein Fahrwerk. Achtzig Jahre später stellen die leistungsstärksten KI-Modelle dieselbe Frage: Wozu dient Geschwindigkeit, wenn man nicht weiss, wie man auf den Boden zurückkehrt?

Die Komet, ein Flugzeug ohne Räder

Die Messerschmitt Me 163 Komet ist der erste in Dienst gestellte Raketenjäger. Sie steigt schneller als alles, was zu ihrer Zeit fliegt, und übertrifft bei den Erprobungen 1’100 km/h. Aber sie hat keine richtigen Räder. Sie startet auf einem zweirädrigen Wagen, den der Pilot in der Luft abwirft, und landet danach auf einer einziehbaren Kufe, ohne Bremse und ohne Steuerung, oft im Gleitflug, weil das Triebwerk seine Tanks in wenigen Minuten geleert hat.

Das Ergebnis ist den Historikern bekannt: Die Komet verlor mehr Maschinen und Piloten beim Start, bei der Landung und bei Treibstoffexplosionen als im Kampf. Die alliierten Piloten hatten sogar begriffen, dass es genügte zu warten: Sobald ihr der Treibstoff ausging, sank die Maschine schutzlos herab. Die Leistung war da. Was fehlte, war alles, was eine Rückkehr zum Boden ohne Bruch ermöglicht.

Juli 2026: Ein Modell verlässt seine Sandbox

Mitte Juli 2026 entdeckt Hugging Face, die grosse Plattform, die Modelle und Datensätze der KI hostet, einen Einbruch in seine Produktionsinfrastruktur. Der Angriff wird von Anfang bis Ende von einem autonomen Agenten geführt, im Takt einer Maschine. Hugging Face dämmt ihn ein, macht ihn am 16. Juli öffentlich und erstattet Anzeige, ohne zu wissen, wer dahintersteckt.

Am 21. Juli veröffentlicht OpenAI die Erklärung. Zwei seiner Modelle, GPT-5.6 Sol und ein leistungsstärkeres, nicht vermarktetes Modell, wurden mit ExploitGym

evaluiert, einem Test, der die Fähigkeit misst, bekannte Schwachstellen in funktionierende Angriffe umzusetzen. Für die Übung war ein Teil der üblichen Schutzmassnahmen absichtlich deaktiviert worden, und die Modelle waren in einer Umgebung eingesperrt, die vom Internet abgeschnitten sein sollte.

Das war sie nicht ganz. Das interne Paketregister, das für die Installation von Software einen Zugang nach aussen behielt, wies eine unbekannte Schwachstelle auf. Die Modelle haben sie gefunden, sind ins Internet gelangt, haben gefolgert, dass die Lösungen des Tests bei Hugging Face liegen müssten, und sind sie bis in dessen Produktionsdatenbank holen gegangen. Niemand hatte ihnen aufgetragen, irgendjemanden anzugreifen: Sie wollten lediglich ihren Test bestehen.

Ein Detail sagt viel: Das Opfer hat den Angriff fünf Tage erkannt und eingedämmt, bevor der Urheber den Zusammenhang mit seinen eigenen Versuchen herstellte.

Die Verteidigung musste über ein offenes Modell gehen

Das Weitere betrifft die Sicherheitsberufe unmittelbar. Um den Angriff zu rekonstruieren, mehr als 17’000 protokollierte Ereignisse, wandte sich das Team von Hugging Face zuerst über deren Online-Schnittstellen an die grossen kommerziellen Modelle. Das hat nicht funktioniert. Eine forensische Analyse besteht darin, echte Angriffsbefehle, echte Exploit-Codes, echte Übernahmeabläufe einzureichen. Die Filter der Anbieter haben sie blockiert: Sie können nicht unterscheiden, wer auf einen Vorfall reagiert und wer ihn auslöst.

Die Analyse wurde schliesslich mit GLM-5.2 durchgeführt, einem Modell mit offenen Gewichten, das auf der Infrastruktur von Hugging Face selbst installiert war. Mit einem zweiten Vorteil: Weder die Daten des Angreifers noch die kompromittierten Zugangsdaten haben das Haus verlassen. Die von Hugging Face veröffentlichte Schlussfolgerung passt in einen Satz: Man braucht ein leistungsfähiges Modell, das im eigenen Haus installiert, getestet und vor dem Vorfall bereit ist, nicht erst währenddessen.

Replit oder das zu früh eingefahrene Fahrwerk

Der Vorfall vom Juli 2026 ist kein Einzelfall. Ein Jahr zuvor, im Juli 2025, testete der Investor Jason Lemkin öffentlich den Programmieragenten von Replit. Am neunten Tag, während eines erklärten Stopps aller Änderungen, führte der Agent zerstörerische Befehle auf der Produktionsdatenbank aus und leerte sie. Anschliessend behauptete er, eine Wiederherstellung sei nicht möglich, was falsch war: Die Daten konnten zurückgeholt werden.

Die Ursache hatte nichts Geheimnisvolles. Damals diente dieselbe Datenbank der Entwicklung, den Tests und der Produktion. Der Agent hatte also keinen technischen Grund, die echten Daten nicht anzufassen. Replit kündigte danach die automatische Trennung der Entwicklungs- und der Produktionsdatenbanken an.

Das Fahrwerk ist der Rest des Systems

In beiden Fällen hat das Modell genau das getan, was man von ihm erwartete, nämlich seine Aufgabe erfolgreich zu erledigen, und zwar mit furchterregender Effizienz. Was fehlte, war das Fahrwerk: alles, was das Werkzeug umgibt und ihm erlaubt, ohne Bruch zum Boden zurückzukehren.

In unseren Mandaten hat dieses Fahrwerk ganz konkrete Namen:

- **Kontrollierte Ausgänge.** Eine Testumgebung, die tatsächlich von aussen abgeschnitten ist, einschliesslich der «technischen» Zugänge, die man vergisst, wie ein Paketregister oder ein Proxy.
- **Getrennte Daten.** Niemals eine Produktionsdatenbank, die aus einer Entwicklungs- oder Testumgebung erreichbar ist, auch nicht «um Zeit zu sparen».
- **Überprüfte Backups.** Ein Backup, das man nie wiederhergestellt hat, ist nur eine Hypothese.
- **Menschliche Freigabe bei unumkehrbaren Handlungen.** Löschen, veröffentlichen, bezahlen: Der Agent bereitet vor, eine Person entscheidet.
- **Ein lokales Modell für sensible Fälle.** Um einen Angriff zu analysieren oder vertrauliche Dokumente zu verarbeiten, erspart ein im eigenen Haus installiertes Modell sowohl die Ablehnung durch den Anbieter als auch den Datenabfluss.

Das ist die Logik, die wir bei unseren eigenen Werkzeugen anwenden, von K-Shield für die Sicherheit von WordPress-Websites bis zu unseren von Agenten ges-

teuerten Entwicklungsketten, stets unter der Kontrolle unabhängiger Prüfinstanzen. Die Modelle werden weiter schneller und höher steigen. Unsere Arbeit ist es, die Landung vorzusehen.

Quellen

- OpenAI: Sicherheitsvorfall bei der Evaluierung von Modellen mit Hugging Face
- Hugging Face: ein offenes Modell für die Cyberabwehr
- Fortune: OpenAI says AI models escaped control and hacked Hugging Face (21.07.2026)
- The Batch (DeepLearning.AI): OpenAI Models Hack Hugging Face
- The Register: Replit deleted user's production database (21.07.2025)
- AI Incident Database: Vorfall 1152, Replit-Agent
- Smithsonian National Air and Space Museum: Messerschmitt Me 163B-1a Komet