Lakera’s 20M-Paradox: Warum das Zürcher Startup mit “Gandalf” eine Illusion der KI-Sicherheit verkauft

Posted on August 12, 2025

AKTE-AI-250812-180: Ein Zürcher Startup sichert 20 Millionen Dollar für KI-Sicherheit mit einem Spiel, das beweist, wie unsicher KI ist. Die Ironie dahinter offenbart die grösste Schwachstelle der AI-Security-Branche.

Das Gandalf-Paradoxon: Wenn Sicherheitstests zur Geschäftsgrundlage werden

Im Juli 2024 verkündete das Zürcher Startup Lakera stolz eine Series-A-Finanzierung von 20 Millionen US-Dollar. Lead-Investor Atomico und andere Geldgeber investierten in eine Vision: die Absicherung von KI-Systemen gegen Manipulation. Was sie tatsächlich finanzierten, ist jedoch ein faszinierendes Paradoxon der modernen Tech-Welt.

Lakera’s Kernprodukt basiert auf “Gandalf” – einem viralen Online-Spiel, bei dem über

500.000 Nutzer versuchen, eine KI zu überlisten und ihr ein geheimes Passwort zu entlocken. Die Spieler generieren dabei täglich zehntausende neue Angriffsmuster. Diese Daten – über 50 Millionen Datenpunkte – bilden das Fundament von Lakera’s Sicherheitstechnologie.

> Ein Sicherheitsunternehmen, dessen Geschäftsmodell darauf basiert, dass Menschen beweisen, wie leicht KI-Systeme zu manipulieren sind – und das dafür 20 Millionen Dollar erhält.

Die Mechanik der Manipulation: Was Gandalf wirklich offenbart

Das Gandalf-Spiel funktioniert nach einem simplen Prinzip: Eine KI hütet ein Passwort, und Spieler versuchen durch geschickte Prompt-Formulierungen, dieses Geheimnis zu extrahieren. Mit jedem Level steigen die Sicherheitsmassnahmen – theoretisch. In der Praxis zeigt das Spiel jedoch etwas anderes:

- Level 1-3: Einfache direkte Fragen reichen aus
- Level 4-5: Umformulierungen und indirekte Ansätze funktionieren
- Level 6-7: Kreative Workarounds überlisten selbst “gehärtete” Systeme

Jeder erfolgreiche Hack wird von Lakera analysiert und in ihre Sicherheitslösung integriert. Doch hier liegt der fundamentale Widerspruch: Je mehr Menschen das System knacken, desto mehr “Sicherheitsdaten” generiert Lakera – und desto deutlicher wird, dass keine KI wirklich sicher ist.

Die Zahlen sprechen eine deutliche Sprache

Metrik	Wert	Bedeutung
Spieler	500.000+	Potenzielle Angreifer im echten Leben
Tägliche neue Angriffe	10.000+	Ständig neue Schwachstellen
Erfasste Datenpunkte	50+ Millionen	Beweis der Vielfalt möglicher Angriffe
Erfolgsrate	Nicht publiziert	Vermutlich zu hoch für Komfort

Der Schweizer Kontext: Zwischen ETH-Innovation und Venture-Capital-Realität

Lakera positioniert sich geschickt im Schweizer AI-Ökosystem. Als Zürcher Startup profitiert es von der Nähe zur ETH und dem wachsenden Ruf der Schweiz als AI-Hub. Die Wahl von Atomico als europäischem Lead-Investor ist dabei kein Zufall – sie spiegelt einen Trend wider:

1. Schweizer AI-Startups suchen vermehrt europäische statt US-amerikanische Finanzierung
2. Gleichzeitig planen sie die Expansion in den US-Markt
3. Die "Swiss Made"-Qualität wird als Verkaufsargument für Sicherheitslösungen genutzt

Lakera plant, das Team von 30 auf 60 Mitarbeiter zu verdoppeln und ein US-Büro zu eröffnen. Die Ironie: Sie exportieren eine Sicherheitslösung, die auf der dokumentierten Unsicherheit von KI basiert.

Die Kundenliste: Wenn Grossunternehmen auf Illusionen setzen

Microsoft, Dropbox und andere Grossunternehmen nutzen bereits Lakera's API-basierte Sicherheitslösung. Diese Unternehmen investieren Millionen in eine Technologie, die bestenfalls ein Moving Target verteidigt. Die Realität sieht so aus:

- **Microsoft**: Integriert Lakera-Technologie, während eigene KI-Systeme regelmässig kompromittiert werden
- **Dropbox**: Setzt auf Lakera für KI-Sicherheit, obwohl das Grundproblem ungelöst bleibt
- **Weitere Enterprise-Kunden**: Kaufen sich ein Gefühl von Sicherheit, keine echte Sicherheit

Das 80-Prozent-Problem

Schätzungen zufolge werden 80% der Unternehmen in naher Zukunft GenAI einsetzen. Diese Statistik wird von Lakera als Marktchance präsentiert. Tatsächlich offenbart sie ein systemisches Problem:

> Wenn 80% der Unternehmen auf eine Technologie setzen, die fundamental unsicher ist, und die beste verfügbare “Sicherheitslösung” auf einem Spiel basiert, das diese Unsicherheit demonstriert – wo führt das hin?

Die technische Realität hinter dem Marketing

Lakera’s Ansatz basiert auf Pattern Recognition und Machine Learning. Die aus Gandalf gewonnenen Angriffsmuster werden verwendet, um potenzielle Prompt Injections zu erkennen. Das klingt technisch solide, ignoriert aber fundamentale Probleme:

Problem 1: Der Wettrüsten-Effekt

Jede neue Sicherheitsmassnahme motiviert kreativere Angriffe. Gandalf selbst beweist dies – mit jedem Level werden die Angriffe sophistizierter, nicht die Verteidigung stärker.

Problem 2: Die Kontextualität von Sprache

Prompt Injection nutzt die Mehrdeutigkeit natürlicher Sprache. Eine Sicherheitslösung, die auf Pattern Matching basiert, kann niemals alle Kontexte und Bedeutungsebenen erfassen.

Problem 3: Der Grundwiderspruch von LLMs

Large Language Models sind darauf trainiert, hilfreich zu sein und Anfragen zu beantworten. Sicherheitsmassnahmen widersprechen diesem Grundprinzip – ein unlösbarer Konflikt.

Die Investoren-Perspektive: Warum 20 Millionen in eine Illusion fliessen

Atomico und andere Investoren sind keine Anfänger. Warum investieren sie in ein Unternehmen, dessen Produkt die eigene Obsoleszenz beweist? Die Antwort liegt in der Natur des Venture Capital:

1. **Market Timing**: Der AI-Security-Markt explodiert, unabhängig von der Effektivität der Lösungen
2. **Exit-Strategie**: Ein Verkauf an einen Tech-Giganten ist wahrscheinlicher als langfristiger Erfolg

3. **Narrative über Realität**: Die Geschichte “Schweizer Qualität sichert KI” verkauft sich besser als technische Wahrheiten

Die Bewertungslogik

Eine 20-Millionen-Series-A impliziert eine Post-Money-Bewertung im dreistelligen Millionenbereich. Für ein Unternehmen, dessen Kernprodukt ein virales Spiel ist, das die Unmöglichkeit seiner Mission demonstriert. Die Bewertung basiert auf:

- Nutzerwachstum (500.000+ Gandalf-Spieler)
- Enterprise-Kunden (Microsoft, Dropbox)
- Marktpotenzial (80% GenAI-Adoption)
- Team und Technologie (30 Mitarbeiter, ETH-Nähe)

Nirgends in dieser Gleichung taucht die Frage auf: Funktioniert es wirklich?

Die Zukunft der KI-Sicherheit: Zwischen Hoffnung und Hybris

Lakera’s Plan, in die USA zu expandieren und das Team zu verdoppeln, zeigt Ambition. Doch die fundamentalen Fragen bleiben:

> Kann man ein System sichern, das per Design darauf ausgelegt ist, auf menschliche Eingaben zu reagieren und dabei “hilfreich” zu sein?

Die Antwort, die Gandalf täglich tausendfach liefert, ist eindeutig: Nein. Dennoch wird die Illusion der Sicherheit teuer verkauft und gekauft.

Alternative Ansätze

Statt auf Pattern Recognition zu setzen, könnten echte Lösungen anders aussehen:

- **Architektonische Isolation**: KI-Systeme von kritischen Funktionen trennen
- **Capability Limitation**: LLMs grundsätzlich einschränken, statt nachträglich absichern
- **Human-in-the-Loop**: Kritische Entscheidungen niemals automatisieren

Diese Ansätze sind weniger sexy, generieren keine viralen Spiele und keine 20-Millionen-Finanzierungen. Aber sie könnten tatsächlich funktionieren.

Das Schweizer Paradox: Präzision trifft auf Prompt Injection

Die Schweiz, bekannt für Präzision und Zuverlässigkeit, beherbergt nun ein Startup, das die Unpräzision und Unzuverlässigkeit von KI monetarisiert. Lakera ist dabei kein Einzelfall – es repräsentiert einen grösseren Trend:

1. Schweizer Startups adaptieren Silicon-Valley-Methoden
2. Wachstum und Narrative werden über Substanz gestellt
3. Der “Swiss Made”-Stempel wird zur Marketing-Floskel

Schlussfolgerung: Die 20-Millionen-Dollar-Lektion

Lakera’s Erfolg ist kein Beweis für die Lösbarkeit des KI-Sicherheitsproblems. Es ist ein Symptom einer Industrie, die bereit ist, Millionen in Lösungen zu investieren, die ihre eigene Unwirksamkeit demonstrieren.

Gandalf, das virale Spiel, ist dabei mehr als nur ein Datensammler – es ist ein Spiegel, der uns zeigt, was wir nicht sehen wollen: **Die aktuelle Generation von KI-Systemen ist fundamental unsicher, und keine Menge von Pattern Recognition oder Schweizer Engineering wird das ändern.**

Die 20 Millionen Dollar, die in Lakera fliessen, sind letztendlich eine Wette darauf, dass die Illusion von Sicherheit profitabler ist als die Konfrontation mit der Realität. In einer Welt, in der 80% der Unternehmen auf GenAI setzen werden, mag das sogar stimmen.

Aber Gandalf, der digitale Zauberer, der sein Geheimnis nicht bewahren kann, erinnert uns daran: **Manchmal ist die grösste Magie die Fähigkeit, anderen glauben zu machen, dass Magie real ist – auch wenn sie es nicht ist.**