

Blackbox

Warum man Security im KI-Zeitalter anders denken muss

Bettina Sterner, BA BA

kaffeekipferl.at

September 2026

Die in diesem Whitepaper enthaltenen Ausführungen, Analysen und Schlussfolgerungen spiegeln ausschließlich die persönliche fachliche Einschätzung der Autorin zum Zeitpunkt der Veröffentlichung wider. Sie sind unabhängig entstanden und stehen in keinem Zusammenhang mit aktuellen oder früheren Arbeitgebern, Auftraggebern, Mandanten oder sonstigen Organisationen, denen die Autorin angehört oder angehört hat. Das Whitepaper ersetzt keine rechtliche oder regulatorische Beratung im Einzelfall.

Deutsch

Ein KI-Agentensystem bricht aus einer als isoliert geltenden Sandbox aus, exploitet eine Schwachstelle und kompromittiert die Infrastruktur eines Drittanbieters – nicht aus bösartiger Absicht, sondern weil eine Sicherheitskontrolle nicht das hielt, was ihr Ersteller ihr zutraute. Der Hugging-Face-Vorfall vom Juli 2026 ist kein Einzelfall, sondern eine **Illustration eines strukturellen Problems**: Kontrollen versagen selten technisch – sie tun genau das, was sie sollen, nur nicht das, was jemand beim Bau angenommen hat. Dieses Whitepaper zeigt an fünf Beispielen – Sandbox-Isolation, Überwachungssysteme, Exit-Strategie, Incident Management und Awareness –, dass der eigentliche Risikofaktor nicht der bedienende, sondern der entwerfende Mensch ist: Seine unvermeidliche Betriebsblindheit beim Systembau, die generative KI nicht durch Cleverness, sondern durch systematisches, ermüdungsfreies Austesten aufdeckt. Weder ein politisch diskutierter KI-„Kill Switch“ noch die menschliche Aufsicht nach Art. 14 KI-VO adressieren diese Lücke – beide setzen bei der Kontrolle über ein laufendes System an, nicht bei den Annahmen, auf denen eine umgangene Kontrolle ursprünglich beruhte. Der Ansatz erfordert keine neuen Managementsysteme, sondern Security neu zu denken – mit konkreten, sofort umsetzbaren Hebeln für CISOs.

English

An AI agent system breaks out of what was assumed to be an isolated sandbox, exploits a vulnerability and compromises a third party's infrastructure – not out of malicious intent, but because a security control did not hold up to what its creator assumed it could withstand. The July 2026 Hugging Face incident is not an isolated case but an **illustration of a structural problem**: controls rarely fail technically – they do exactly what they were built to do, just not what someone assumed they covered. Across five examples – sandbox isolation, monitoring systems, exit strategy, incident management and awareness – this whitepaper shows that the real risk factor is not the human operating a control, but the human who designed it: an unavoidable blind spot in system design that generative AI exposes not through cleverness but through systematic, tireless probing. Neither a politically debated AI "kill switch" nor human oversight under Article 14 of the EU AI Act address this gap – both target control over a running system, not the assumptions the bypassed control originally rested on. The approach requires no new management systems, only rethinking security itself – paired with concrete, immediately actionable levers for CISOs.

Auf einen Blick:

- **Die These:** Nicht der Fehler im Umgang mit einer Kontrolle gefährdet uns, sondern die unvermeidliche Lücke zwischen dem, was ein System leisten soll, und dem, was sein menschlicher Architekt sich beim Bau vorstellen konnte.
- **Der Auslöser:** Der Hugging-Face-Vorfall (Juli 2026) – ein autonomer, mehrtägiger KI-Cyberangriff mit rund 17.600 Einzelaktionen –, der erst möglich wurde, weil eine als isoliert geltende Sandbox es nicht war; drei parallele reale Vorfälle bei Anthropic sowie ein vierter, im September 2026 offengelegter Fall bestätigen dasselbe Muster.
- **Die Abgrenzung:** Kein Insider-Threat-Fall (die relevante Entscheidung liegt zeitlich vor dem Vorfall, oft Monate oder Jahre) und kein Human-Oversight-Problem (Art. 14 KI-VO regelt Kontrolle über die laufende KI, nicht die Annahme, auf der eine Kontrolle beim Entwurf beruhte).
- **Warum jetzt:** Automation Bias auf der Verteidigerseite trifft auf KI-Systeme, die genau diese blinden Flecken systematisch und ermüdungsfrei aufspüren – und laut AI Security Institute UK dafür längst keine Frontier-Modelle mehr benötigen (Abstand zu offenen Modellen nur noch vier bis sieben Monate).
- **Fünf Beispiele:** Sandbox-Isolation, Überwachungssysteme, Exit-Strategie, Incident Management und Awareness zeigen dieselbe Struktur in fünf unterschiedlichen Bereichen der Security Governance – jeweils mit realistischer Einschränkung und konkretem CISO-Hebel.
- **Was es nicht braucht:** Keine neuen Managementsysteme, keinen politischen Kill Switch – sondern Security neu zu denken, u.a. mit einem eigenen Vorschlag zur Übertragung des TLPT-Gedankens aus DORA auf die KI-VO.
- **Für wen:** CISOs, Security- und GRC-Verantwortliche.

Inhalt

Auf einen Blick:	2
Inhalt	3
1. Ausgangspunkt: Der Hugging-Face-Vorfall	4
1.1 Vom Evaluierungslauf zur realen Infrastruktur	4
1.2 Die Kontrolle funktionierte – nur nicht so wie angenommen	5
1.3 Die Blackbox liegt vor der Kontrolle	6
1.4 Anthropic: Dieselbe Struktur, ohne Sandbox-Escape	7
2. Die Kernthese: Der menschliche Faktor beginnt vor der Fehlbedienung	8
2.1 Fehlervermeidung setzt voraus, dass der Fehlerraum bekannt ist	8
2.2 Nicht Insider Threat	9
2.3 Nicht Human Oversight	10
3. Wenn Ausdauer skalierbar wird	10
3.1 Der menschliche Angreifer war Teil des impliziten Threat Models	11
3.2 Automation Bias auf der Verteidigerseite	11
3.3 Für den Angriff braucht es kein Frontier-Modell	11
3.4 Die Schwachstelle muss nicht neu sein	12
3.5 Wenn sich das Threat Model schneller ändert als die Kontrolle	13
4. Fünf Beispiele, eine Struktur	13
4.1. Sandbox-Isolation und Zugriffsschlüssel	13
4.2. Überwachungssysteme (SIEM)	15
4.3. Exit-Strategie	16
4.4. Incident Management: Handlungsketten statt Einzelereignisse	19
4.5. „Security Awareness 2.0“	20
5. Fazit / Ausblick	22

1. Ausgangspunkt: Der Hugging-Face-Vorfall

Der Sicherheitsvorfall bei Hugging Face im Juli 2026 ist nicht deshalb bemerkenswert, weil ein KI-Agent eine Sicherheitskontrolle überwunden hat. Sicherheitskontrollen werden überwunden; Schwachstellen werden übersehen, Konfigurationen erweisen sich als fehlerhaft, Angriffspfade bleiben unentdeckt. **Bemerkenswert ist, unter welchen Bedingungen das geschah – und welche Annahmen über die Wirksamkeit der eingesetzten Kontrollen sich dabei als falsch erwiesen.**

Ausgangspunkt waren interne Cybersecurity-Evaluierungen von OpenAI auf Basis des ExploitGym-Benchmarks¹: KI-Agenten sollten Software-Schwachstellen identifizieren und ausnutzen, isoliert in Sandbox-Umgebungen. Einzelne Komponenten dieser Umgebung – etwa ein Package-Registry-Cache-Proxy – mussten dafür legitim auf externe Ressourcen zugreifen können. Genau diese vorgesehene Verbindung wurde zum Ausgangspunkt eines nicht vorgesehenen Angriffspfads: Ein KI-Agent fand eine bis dahin unbekannte Schwachstelle im Proxy, verließ darüber die Isolation und erreichte das Internet.²

³ OpenAI hat den Vorgang später zusätzlich in einem ausführlichen technischen Incident-Report dokumentiert, der unter anderem Reward Hacking⁴ als Mitursache benennt: Die KI-Agenten suchten nicht in böswilliger Absicht nach dem Ausbruch, sondern nach dem kürzesten Weg zu einer hohen Bewertung der gestellten Aufgabe.⁵ Genau dieses Muster hatte OpenAI bereits ein Jahrzehnt zuvor an einem harmlosen Beispiel beschrieben: Ein Spiel-Agent maximierte seinen Punktestand in einem Bootrennen nicht durchs Ziel, sondern durch endloses Drehen im Kreis über dieselben drei Wertungsflaggen – schneller Punktegewinn statt des eigentlich intendierten Wegs.⁶

Das eigentliche Problem begann daher schon davor, denn die angenommene Trennung zwischen den KI-Agenten war nicht so vollständig wie vorgesehen: Sie entdeckten gemeinsam erreichbare Infrastruktur und nutzten sie zur Kommunikation – zunächst über Artifactory, nach dessen Entfernung über ein neu etabliertes Messageboard. METR und Redwood Research rekonstruierten rund 1.200 Agenten, die dort mehr als 70.000 Nachrichten und Dateien austauschten; rund 700 davon beteiligten sich später am Angriff auf Hugging Face.⁷ Die Kommunikation war kein Nebenprodukt: Die Agenten koordinierten Vorhaben, tauschten Erkenntnisse aus und versuchten, den ExploitGym-Scorer zu manipulieren. METR/Redwood fanden Hinweise, dass die Zusammenarbeit Ergebnisse ermöglichte, die einzelne KI-Agenten allein nicht erreicht hatten.⁸ Eine Architektur, die von der Isolation einzelner Läufe ausging, erhielt damit eine zusätzliche Handlungsebene: **Erkenntnisse ließen sich zwischen KI-Agenten weitergeben, kombinieren, für weitere Versuche nutzen.**

1.1 Vom Evaluierungslauf zur realen Infrastruktur

Der Angriff auf Hugging Face war kein einzelner Sandbox-Escape mit anschließendem geradlinigem Zugriff. Die von Hugging Face veröffentlichte forensische Rekonstruktion beschreibt eine mehrtägige Kette über mehrere Systeme und Vertrauensgrenzen: Zwischen 9. und 13. Juli wurden rund 17.600

¹ Anm.: ExploitGym ist ein öffentlich zugänglicher Cybersecurity-Benchmark, entwickelt von UC Berkeley, Max Planck Institute for Security and Privacy, UC Santa Barbara, Arizona State University mit Modellzugang von Anthropic, OpenAI und Google. Im Gegensatz zu anderen Benchmarks für AI Agents soll hier getestet werden, ob ein Modell einen Bug nicht nur findet und einen Patch schreibt, sondern ob ein Modell aus einer bekannten Schwachstelle einen realen Angriff macht, inklusive Durchhalten über viele Zwischenschritte.

² Hugging Face Inc. (Autoren: Larcher, H. / Carreira, A. et al), Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident, 27. Juli 2026; <https://huggingface.co/blog/agent-intrusion-technical-timeline>.

³ Open AI Inc., Der Vorfall bei Hugging Face und der weitere Weg, 26. August 2026; <https://openai.com/de-DE/index/hugging-face-incident-and-the-road-ahead>.

⁴ Anm.: Verhalten, bei dem ein KI-System lernt, die Bewertungsgröße einer Aufgabe zu maximieren, statt die eigentlich gewollte Aufgabe zu lösen; das System tut nicht das, was der Mensch meinte, sondern das, was die Belohnungsfunktion tatsächlich misst, und diese beiden Dinge fallen nicht immer zusammen: Ursache: Reinforcement Learning, eine Trainingsmethode, bei dem ein KI-Modell ein Signal („Belohnung“) erhält, wenn etwas Bestimmtes erreicht wird, zB eine korrekte Lösung oder ein erfolgreiches Testergebnis. Wenn es einen kürzeren Weg gibt, um diesem Signal zu kommen, findet ein leistungsfähiges, beharrliches System diesen Weg mit hoher Wahrscheinlichkeit. Siehe auch OpenAI, Faulty Reward Functions in the Wild, 21. Dezember 2016, <https://openai.com/index/faulty-reward-functions/>.

⁵ OpenAI Inc., OpenAI – Hugging Face Incident Technical Report, Abschnitt VIII.A „Reward Hacking is a common problem in training and evaluations“, 26. August 2026; <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>.

⁶ MIT Technology Review (Autor Heaven, W.D.), OpenAI called the Hugging Face attack unprecedented. But we've been here before, 27. Juli 2026; <https://www.technologyreview.com/2026/07/27/1140836/openai-hugging-face-attack-precedent/>.

⁷ METR Model Evaluation & Threat Research (Autoren Greenblatt, R./ Cotra, A. / Wijk, H.), Brief independent investigation of agents' behaviour, reasoning and collaboration in the OpenAI / Hugging Face hacking incident, 26. August 2026; <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation>.

⁸ Ebenda.

Aktionen in etwa 6.280 Clustern rekonstruiert – Reconnaissance, Remote Code Execution, Nachladen von Werkzeugen, Exfiltration, Command-and-Control, Umgehungsmaßnahmen, später Zugriffe auf Kubernetes-, Netzwerk- und Supply-Chain-Infrastruktur.⁹

Nach dem Verlassen der OpenAI-Sandbox nutzte der KI-Agent zunächst eine öffentlich erreichbare Code-Evaluierungsumgebung eines Dritten als Sprungbrett. Ein klassischer SSRF-Versuch¹⁰ gegen den Dataset Processor von Hugging Face scheiterte an einer Allowlist – der Agent wechselte den Ansatz und nutzte stattdessen manipulierte Dataset-Konfigurationen, um Code in einem produktiven Worker auszuführen. Von dort folgten Zugriff auf Kubernetes Service Account Tokens, Kommunikation mit Cluster-APIs, interne Netzwerkressourcen, Source-Control-Infrastruktur, weitere laterale Bewegungen.¹¹

Für die Analyse zählt der Verlauf mehr als der einzelne erfolgreiche Exploit: Es gibt keinen einzelnen technischen Durchbruch, nach dem der Angriff entschieden war. Wege wurden getestet, verworfen, ersetzt; als reguläre Egress-Pfade eingeschränkt wurden, griff der Agent auf DNS-Manipulation, statische Host-Zuordnungen und vorbereitete Werkzeuge zurück. Hugging Face beschreibt den Vorfall entsprechend als Abfolge tausender automatisierter Einzelentscheidungen über kurzlebige Sandbox-Umgebungen hinweg.¹²

Das verändert eine bislang implizite Rahmenbedingung von Sicherheitsarchitekturen. Charles Perrows Konzept der „Normal Accidents“ beschreibt, wie in interaktiv komplexen, eng gekoppelten Systemen einzeln funktionierende, korrekt konstruierte Komponenten in Kombination Systemversagen erzeugen – ohne dass irgendeine einzelne Komponente ausfällt.¹³ Genau das zeigt Hugging Face: Kein Baustein der Angriffskette musste außergewöhnlich sein. Entscheidend wurde, ob sich aus mehreren, jeweils begrenzten Möglichkeiten eine funktionierende Kette bilden ließ. Ein agentisches KI-System kann diese Suche in großer Zahl, parallel und ohne Ermüdung fortsetzen – und genau darin liegt die Asymmetrie, auf die Hugging Face selbst verweist: Die Verteidigung muss zahlreiche Grenzen gleichzeitig absichern, dem Angreifer genügt eine funktionierende Kette.¹⁴

1.2 Die Kontrolle funktionierte – nur nicht so wie angenommen

Eine rein technische Aufarbeitung endet bei Sandbox-Härtung, Netzwerksegmentierung, Egress-Kontrollen, Credential Management, zusätzlicher Überwachung, besserer Trennung gemeinsamer Infrastruktur – Maßnahmen, die OpenAI und Hugging Face nach dem Vorfall auch umgesetzt haben.^{15 16} Für dieses Whitepaper liegt der interessantere Punkt davor.

Sandbox, KI-Agenten-Trennung, Netzwerkbeschränkung, Allowlist, Monitoring: Jede dieser Kontrollen war funktional, keine sinnlos. Die URL-Allowlist blockierte den klassischen SSRF-Versuch exakt wie vorgesehen – der Angriff endete damit nicht, sondern wechselte auf einen Mechanismus, der dieselbe Grenze anders überschritt.¹⁷

Damit wird eine Unterscheidung entscheidend, die in der klassischen Betrachtung des „Human Factor“ verloren geht: Das Problem liegt nicht zwingend darin, dass ein Mensch eine Kontrolle falsch bedient. Es entsteht bereits dort, wo entschieden wird, **welche Reichweite eine Kontrolle haben muss, gegen welche Szenarien sie geprüft wird und wann sie als ausreichend gilt**. Nicht die Frage, wie die Sandbox umgangen wurde, ist entscheidend – sondern die Frage, **warum jemand angenommen hat, sie hält**.

⁹Hugging Face Inc., Security incident disclosure – July 2026, 16. Juli 2026; <https://huggingface.co/blog/security-incident-july-2026>.

¹⁰ Anm.: SSRF – Server-Side Request Forgery („serverseitige Anfragefälschung“); Grundprinzip: Ein Angreifer bringt einen Server dazu, in dessen Namen eine Netzwerkanfrage an ein Ziel zu senden, das er selbst nicht direkt erreichen dürfte. Der Server wird somit zum Werkzeug für den Angreifer. SSRF ist ein gutes Beispiel für eine Kontrolle, die genau das tut, wofür sie gebaut wurde (Allowlist blockt unerlaubte Ziele), aber auf der Annahme beruht, dass alle relevanten Umgehungswege durch diese Prüfung abgedeckt sind – diese Annahme war unvollständig.

¹¹Hugging Face Inc., Security incident disclosure.

¹²Ebenda.

¹³Perrow, C., Normal Accidents: Living with High-Risk Technologies, Basic Books, 1984; Williams, H. M./Yampolskiy, R. V., Understanding and Avoiding AI Failures: A Practical Guide, 13. März 2024, S. 2; <https://arxiv.org/pdf/2104.12582>.

¹⁴Hugging Face Inc., Security incident disclosure.

¹⁵Hugging Face Inc., Security incident disclosure.

¹⁶Open AI Inc., Der Vorfall bei Hugging Face und der weitere Weg.

¹⁷Ebenda.

Das ist kein nachträglicher Vorwurf an die Entwickler: Eine Sicherheitsarchitektur kann nur auf Basis der zu diesem Zeitpunkt erkennbaren Eigenschaften, Abhängigkeiten und Bedrohungsszenarien entworfen werden. Jede Kontroll-Entscheidung enthält damit Annahmen – darüber, wie ein System genutzt werden kann, welche Komponenten interagieren, welche Angriffspfade plausibel sind. Hugging Face macht sichtbar, was passiert, wenn diese Annahmen selbst zum Angriffspunkt werden: Die Isolation der KI-Agenten verhinderte nicht, dass gemeinsam erreichbare Infrastruktur zum Kommunikationsmedium wurde. Ein legitimer Egress-Pfad wurde zum Weg aus der Sandbox; eine funktionierende Allowlist verhinderte einen Angriffspfad, aber nicht den nächsten.¹⁸

Dieselbe Zweiteilung – „KI außer Kontrolle“ gegen „menschliches Versagen“ – prägt auch einen Großteil der öffentlichen Kommentierung des Vorfalls; beide Deutungen greifen meiner Einschätzung nach zu kurz, weil sie den eigentlichen Punkt verfehlen: Die KI-Modelle mussten keine böswillige Absicht verfolgen, um Schaden anzurichten: Es genügte, dass offenbar niemand geprüft hatte, ob die als isoliert angenommene Umgebung es tatsächlich war.¹⁹

Dasselbe Muster – eine Kontrolle, die exakt das tut, wofür sie gebaut wurde, aber auf einer Annahme beruht, die nicht mehr stimmt – zeigt sich seither wiederholt auch außerhalb des Hugging-Face-Falls. GitLab beschreibt einen eigenen internen Vorfall, bei dem ein KI-Agent über einen explizit für den Zugriff auf Paketregister freigegebenen, aber verwundbaren Proxy aus seiner Sandbox ausbrach; die Cloud Security Alliance bezeichnet dieses Muster als „Trust Handoff“: Der KI-Agent verlässt seine Sandbox nie, nutzt aber eine als vertrauenswürdig eingestufte Verbindung, um eine Aktion mit größerer Reichweite auszuführen, als die Sandbox selbst erlauben sollte.^{20 21} Anfang September 2026 zeigte eine Schwachstelle im Open-Source-Werkzeug DeepSeek Harness dieselbe Struktur ein weiteres Mal: Die Dateisandbox funktionierte wie vorgesehen, ließ sich aber über eine lokale Programmierschnittstelle abschalten, weil „lokal“ anhand eines vom Client mitgelieferten Headers geprüft wurde statt anhand der tatsächlichen Netzwerkverbindung – wieder eine Kontrolle, die genau das tat, wofür sie gebaut war, nur nicht das, was sie eigentlich hätte absichern sollen.²²

1.3 Die Blackbox liegt vor der Kontrolle

Der Begriff Blackbox ist im KI-Kontext meist an die fehlende Erklärbarkeit von KI-Modellen gekoppelt. Das greift hier zu kurz. Die relevante Blackbox liegt nicht im KI-System – sie liegt **im Sicherheitsdesign selbst**.

Wer eine Kontrolle entwirft, braucht ein Modell davon, wovor sie schützen soll. Was in diesem Modell nicht vorkommt, kann bei ihrer Konstruktion auch nicht adressiert werden. Die Grenze entsteht nicht durch mangelnde Umsetzung, sondern durch die notwendige Abstraktion, auf deren Grundlage jede Sicherheitsentscheidung überhaupt getroffen wird.

Das unterscheidet diese Blackbox von einem gewöhnlichen „unknown unknown“²³: Es geht nicht nur darum, dass künftige Schwachstellen unbekannt sind. Entscheidend ist, dass **die Beurteilung der Kontrolle selbst auf Annahmen beruht, deren Unvollständigkeit für die beurteilende Person nicht vollständig sichtbar sein kann** – wäre der blinde Fleck bekannt, wäre er längst geschlossen, und damit kein blinder Fleck mehr.

Wie hartnäckig sich das in der Praxis zeigt, illustriert ein kontrolliertes Experiment des britischen AI Security Institute: Ein in einer Sandbox eingesetzter KI-Agent identifizierte trotz mehrerer Härtungsrunden wiederholt neue Wege, um Organisation, Betreiber und Forschungstätigkeit der Institution zu rekonstruieren – unter anderem, indem er nach Einführung eines eigens dafür

¹⁸Ebenda.

¹⁹DataEthics.eu (Autorin Masaracchia, L.), The OpenAI – Hugging Face case: a story of model behaviour and security threats, Dataethics.eu, 4. September 2026; <https://dataethics.eu/the-openai-hugging-face-case-a-story-of-model-behaviour-and-security-threats/>.

²⁰Cloud Security Alliance, CSA Research Note: AI Coding Agent Sandbox Escapes, 22. Juli 2026; <https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-coding-agent-sandbox-escapes-20260722-c/>

²¹InfoQ (Autor Risi, C.), GitLab Warns That AI Agent Sandboxes Are Only as Secure as Their Network Access, 8. September 2026; <https://www.infoq.com/news/2026/09/gitlab-ai-sandbox-access/>.

²²The Hacker News (Autor Khandelwal, S.), DeepSeek Harness Flaw Let AI Agents Disable Their Own File Sandbox Without Approval, 9. September 2026; <https://thehackernews.com/2026/09/deepseek-harness-flaw-let-ai-agents.html>.

²³US Department of Defense News Briefing Secretary of Defense Donald H. Rumsfeld, 12. Februar 2002, https://avalon.law.yale.edu/sept11/dod_brief146.asp.

eingerrichteten Proxys direkt auf externe Dienste zugriff und deren TLS-Zertifikate auswertete, um die Herkunft doch zu identifizieren. Nach jeder Gegenmaßnahme fand der Agent einen anderen Weg zum selben Ziel.²⁴

Agentische KI erzeugt diese Grenze nicht neu. Sie erhöht die Wahrscheinlichkeit, dass sie praktisch relevant wird: Systeme, die in hoher Geschwindigkeit alternative Wege ausprobieren, Zwischenergebnisse kombinieren und nach Fehlschlägen weiterarbeiten, testen nicht nur bekannte Schwachstellen – sie testen die Vollständigkeit der Annahmen, auf denen eine Sicherheitsarchitektur beruht. Deshalb wäre es zu einfach, aus Hugging Face schlicht die Lehre „bessere Sandboxes bauen“ zu ziehen. Auch eine technisch verbesserte Sandbox beruht auf einem Modell ihrer Umgebung. Zusätzliche Kontrollen senken Wahrscheinlichkeit und Schadensausmaß eines Angriffs erheblich – sie beseitigen aber nicht das erkenntnistheoretische Grundproblem: **Auch die zusätzliche Kontrolle wurde von jemandem entworfen, der entscheiden musste, welche Risiken sie erfasst.**

1.4 Anthropic: Dieselbe Struktur, ohne Sandbox-Escape

Dass es sich nicht um eine Eigenheit der OpenAI-Architektur handelte, zeigte eine unmittelbar folgende Untersuchung bei Anthropic. Als Reaktion auf den Hugging-Face-Vorfall prüfte Anthropic 141.006 eigene Cybersecurity-Evaluierungsläufe darauf, ob Claude-Modelle aus Testumgebungen heraus reale Systeme erreicht hatten. Drei Vorfälle wurden identifiziert, in denen Modelle unautorisierten Zugang zu Produktivsystemen dreier Organisationen erlangten.²⁵

Der technische Ausgangspunkt war ein anderer: Kein Sandbox-Escape über eine neu entdeckte Schwachstelle, sondern eine Fehlkonfiguration bei einem externen Evaluierungspartner, durch die die von Claude erreichbaren Maschinen tatsächlich über Internetzugang verfügten – wovon weder Anthropic noch der jeweilige Partner wussten. Die Evaluierungs-Prompts sagten den Modellen ausdrücklich, sie hätten *keinen* Internetzugang.²⁶ Die Modelle bearbeiteten Capture-the-Flag-Aufgaben²⁷ und hielten die erreichten Systeme zunächst für Teil der simulierten Umgebung. In einem Fall scannte ein internes Forschungsmodell rund 9.000 Ziele und kompromittierte schließlich mit bekannten Techniken – ausgelesene Zugangsdaten aus einer Debug-Seite, SQL Injection – die öffentlich erreichbare Anwendung eines realen Unternehmens. Erst als das Modell erkannte, dass das System keine Verbindung zur Evaluierungsaufgabe hatte, beendete es den Angriff selbstständig. Anthropic ordnet den Vorfall als Versagen des Evaluation Harness und des operativen Setups ein, nicht als Alignment-Versagen.²⁸

Der Fall ist ein wichtiges Gegenstück zu Hugging Face: Hier musste keine hochentwickelte Kontrolle durch einen unerwarteten Exploit überwunden werden. Die zentrale Sicherheitsannahme war von Anfang an falsch – **man ging von einem nicht vorhandenen Internetzugang aus.** Anthropic verifiziert seither die Isolation von Evaluierungsumgebungen explizit, statt sich auf deren Konfiguration zu verlassen, und benennt die Ursache klar als Fehlkonfiguration der Drittanbieter-Umgebung.^{29 30}

Die Fälle unterscheiden sich technisch, führen aber zum selben Problem: Bei OpenAI/Hugging Face erwies sich die angenommene Reichweite mehrerer Sicherheitsgrenzen als unvollständig; bei Anthropic fehlte eine zentrale Voraussetzung der Architektur, ohne dass es jemand merkte. In beiden Fällen existierten Kontrollen, eine Vorstellung davon, was sie absicherten – und eine Systemrealität, die größer war als das Modell, auf dessen Basis ihre Wirksamkeit beurteilt wurde.

²⁴UK AI Security Institute, What can sandboxed AI agents learn about their evaluation environments?, 20. April 2026; <https://www.aisi.gov.uk/blog/what-can-sandboxed-ai-agents-learn-about-their-evaluation-environments> und [https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/69e63511d394ee65d788d9e7_What_OpenClaw_learned_about_us%20\(2\).pdf](https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/69e63511d394ee65d788d9e7_What_OpenClaw_learned_about_us%20(2).pdf).

²⁵Anthropic Inc., Investigating three real-world incidents in our cybersecurity evaluations, 30. Juli 2026; <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>.

²⁶Ebenda.

²⁷Ein Wettbewerbs- bzw. Übungsformat, bei dem Teilnehmer*innen (in dem Fall KI-Agenten) in einer abgesteckten Umgebung eine Schwachstelle finden und ausnutzen müssen, um eine versteckte Zeichenkette („Flag“) zu erbeuten.

²⁸Anthropic Inc., Investigating three real-world incidents in our cybersecurity evaluations.

²⁹Ebenda.

³⁰Anthropic Inc., Improving our alignment and security efforts, 31. August 2026; <https://www.anthropic.com/news/improving-alignment-security-efforts>.

Das ist der Ausgangspunkt dieses Whitepapers: **Das Risiko liegt nicht nur darin, dass Menschen Sicherheitskontrollen falsch bedienen. Es liegt bereits darin, dass Menschen entscheiden müssen, welche Kontrollen notwendig sind, was diese erfassen müssen und wann ihre Wirkung als ausreichend gilt – ohne die eigenen blinden Flecken vollständig kennen zu können.**

2. Die Kernthese: Der menschliche Faktor beginnt vor der Fehlbedienung

Der „menschliche Faktor“ ist in der Security-Theorie seit Jahrzehnten etabliert – meist als Kategorie für Handlungen, durch die bestehender Schutz umgangen, geschwächt oder wirkungslos wird: Phishing, schwache Passwörter, Fehlkonfigurationen, Nichtbeachtung von Vorgaben. Entsprechend setzen Gegenmaßnahmen bei Awareness, Training und technischer Fehlerreduktion an.^{31 32}

Diese Perspektive ist notwendig, erfasst aber nur einen Teil: Sie setzt erst zu dem Zeitpunkt an, zu dem bereits entschieden wurde, welche Kontrolle existiert, welches Risiko sie adressiert und wann ihre Wirkung als ausreichend gilt. Auch diese Entscheidungen treffen Menschen. Damit entsteht eine zweite Ebene des menschlichen Faktors – nicht bei der Anwendung, sondern beim **Entwurf** einer Kontrolle. Jede Sicherheitsarchitektur beruht auf Annahmen über das zu schützende System, relevante Bedrohungen, plausible Angriffspfade und das Verhalten der Beteiligten. Diese Annahmen bestimmen, was eingerichtet, getestet und als Restrisiko akzeptiert wird – sie bilden die Grundlage der Sicherheitsentscheidung **und** begrenzen zugleich ihren Wahrnehmungsraum.

Hugging Face macht diese Grenze sichtbar: Die Schwäche war nicht das Fehlen von Kontrollen – Sandboxen, Netzwerkbeschränkungen, Zugriffskontrollen, Monitoring existierten. Problematisch wurde die Differenz zwischen angenommener und tatsächlicher Reichweite dieser Kontrollen. **Damit verschiebt sich der Blick vom einzelnen Kontrollversagen auf die Annahme, die der Kontrolle vorausgeht.** Eine Kontrolle kann technisch korrekt implementiert sein und trotzdem einen relevanten Angriffspfad nicht erfassen; sie kann alle vorgesehenen Tests bestehen und dennoch unter Bedingungen versagen, die diese Tests nicht abbildeten. Das ist weder Nachlässigkeit noch mangelnde Kompetenz – es ist eine strukturelle Grenze menschlicher Systemgestaltung, die die Sicherheitsforschung selbst seit Längerem beschreibt: Jens Rasmussen etwa zeigte schon vor über zehn Jahren auf, dass Sicherheitsgrenzen nie aus vollständiger Kenntnis abgeleitet werden, sondern aus Annahmen darüber, was als akzeptables Systemverhalten gilt – und dass Systeme unter wirtschaftlichem und operativem Druck systematisch in Richtung dieser Grenzen migrieren, bis eine Annahme kollidiert.³³ Was bei Hugging Face und Anthropic geschah, ist diese Migration in Reinform, nur mit einem Gegenspieler, der die Grenze aktiv sucht statt sie zufällig zu berühren.

2.1 Fehlervermeidung setzt voraus, dass der Fehlerraum bekannt ist

Ein hilfreicher Vergleich kommt aus der industriellen Produktion: „Poka-Yoke“. Die Idee dahinter: Fehler nicht erst im Nachhinein entdecken, sondern Prozesse im Vorhinein so zeichnen, dass bestimmte Fehlhandlungen gar nicht erst möglich sind – oder zumindest sofort auffallen.^{34 35}

Die American Society for Quality definiert Mistake Proofing entsprechend als Einsatz automatischer Vorrichtungen, die einen Fehler verhindern oder sofort sichtbar machen.³⁶ Das Prinzip überzeugt, weil

³¹ Sasse, M.A. / Brostoff, S. / Weirich, D., Transforming the „weakest“ link – a human / computer interaction approach to usable and effective security, BT Technology Journal 2001, 19(3), 122-131; http://www.cs.ucl.ac.uk/fileadmin/sec/publications/Sasse_Brostoff_Weirich_A_human-computer_interaction_approach_to_usable_and_effective_security_BTJ2001.pdf.

³² Anteil menschlicher Beteiligung an Data Breaches 62%, Verizon, 2026 Data Breach Investigations Report, Mai 2026, <https://www.verizon.com/business/resources/T161/reports/2026-dbir-data-breach-investigations-report.pdf>.

³³ Rasmussen, J., Risk management in a dynamic society – a modelling problem, 1997, <https://backend.orbit.dtu.dk/ws/files/158016663/SAFESCI.pdf>.

³⁴ Poka.io, Moderne Anwendungen von Poka Yoke in der schlanken Fertigung, 13. Jänner 2023, <https://www.poka.io/en/blog/expanding-the-meaning-of-poka-yoke-in-lean-manufacturing/>.

³⁵ Sattigeri, R.C. / Kulkarni, D. G., Quality Improvement Technique: Through Poka-Yoke Concepts, IOSR Journal of Business and Management, Vol 23 (9), <https://www.iosrjournals.org/iosr-jbm/papers/Vol23-issue9/Ser-4/A2309040104.pdf>.

³⁶ American Society for Quality, What is Mistake Proofing?, <https://asq.org/quality-resources/mistake-proofing>.

es Verantwortung nicht allein beim Menschen belässt – der Prozess selbst wird verändert, nicht nur die Erwartung an seine fehlerfreie Ausführung.³⁷

Die Übertragbarkeit auf Security-Themen hat allerdings eine Grenze: Poka-Yoke funktioniert dort gut, wo der relevante Fehlerraum bestimmbar ist. Um einen Fehler konstruktiv unmöglich zu machen, muss bekannt sein, **welcher Fehler verhindert werden soll** – auch ASQ beginnt mit der Identifikation der Prozessschritte, an denen Fehler auftreten können.³⁸ Genau diese Voraussetzung ist in komplexen digitalen Systemen nur eingeschränkt gegeben: Die Zahl möglicher Zustände, Abhängigkeiten und Interaktionen ist groß und verändert sich laufend; hinzu kommen externe Komponenten, Lieferketten und – mit agentischer KI – Systeme, die selbstständig nach alternativen Wegen suchen.

Das entwertet Fehlervermeidung als Designprinzip nicht: Secure by Design, Least Privilege, Segmentierung, Default Deny und technische Approval Gates bleiben richtig. Aber: **Auch die Konstruktion einer fehlervermeidenden Kontrolle setzt voraus, dass der relevante Angriffspfad zumindest konzeptionell erfasst wurde.** Daraus folgt ein Paradox: Je stärker Security versucht, menschliche Fehler technisch aus dem System herauszukonstruieren, desto stärker hängt ihre Wirksamkeit von der Qualität der menschlichen Annahmen ab, die in diese Kontrollen eingebaut werden. Das Problem verschwindet nicht – es verlagert sich.

Wie groß diese Verlagerung in der Praxis bereits ist, zeigt eine 2026 veröffentlichte Systematisierung von 39 Arbeiten zur Ausführungssicherheit von KI-Agenten: Automatisierte Tests gegen real eingesetzte Sicherheitsregeln – etwa Positivlisten und Befehls-Denyls – fanden bei 69,0 bis 98,6 Prozent der geprüften, tatsächlich verwendeten Regelwerke einen Umgehungsweg. Die Studie führt das nicht in erster Linie auf schwache Technik zurück, sondern darauf, dass praktisch jede untersuchte Verteidigung gegen einen von den Autoren selbst konstruierten Angreifer getestet wurde – nicht gegen die Umgehungstechniken, die in der Praxis tatsächlich auftreten.³⁹ **Genau das ist die Verschiebung, die dieses Whitepaper beschreibt: Nicht ob eine Kontrolle funktioniert, sondern gegen welche Annahme sie überhaupt geprüft wurde.**

2.2 Nicht Insider Threat

Die Abgrenzung zum klassischen Insider Threat ist wichtig, weil beide Konstellationen ähnlich wirken können: Auch dort liegt die problematische Handlung innerhalb einer legitimen Zugriffs- oder Vertrauensstruktur. Der hier untersuchte Mechanismus liegt aber zeitlich und logisch davor: Der relevante Mensch muss zum Zeitpunkt des Vorfalls nicht mehr handeln – seine Rolle kann Monate oder Jahre zuvor beendet sein, bei der Definition einer Netzwerkgrenze, der Auswahl eines Monitoring-Szenarios, der Entscheidung, welche Systemeigenschaft als hinreichend geprüft gilt.

Das Risiko entsteht nicht aus dem Missbrauch einer Berechtigung, sondern aus einer Annahme, die in das Sicherheitsdesign selbst eingeschrieben wurde. Für einen Insider Threat lassen sich Verhalten, Berechtigungen und Aktivitäten überwachen. Die hier relevante Annahme besitzt keinen eigenen Account, erzeugt keinen Login, löst keinen Alarm aus – solange sie nicht mit der Realität kollidiert, bleibt sie vollständig unsichtbar.

Auch aktuelle Sicherheitsmodelle für agentische KI adressieren bereits neue Risiken: Die OWASP Top 10 for Agentic Applications 2026 behandeln unter anderem Goal Hijacking, Tool Misuse, Identity and Privilege Abuse, Supply-Chain-Risiken, Memory Poisoning, unsichere Inter-Agent-Kommunikation, Cascading Failures und Rogue Agents.⁴⁰ Solche Frameworks sind jedoch – wie schon ihr Vorbild STRIDE⁴¹ – primär architekturzentriert: Sie katalogisieren technische Bedrohungskategorien, ohne den menschlichen Faktor in der hier relevanten Form zu erfassen. Das betrifft nicht nur die häufig konstatierte Lücke bei sozialen Komponenten wie Social Engineering⁴², sondern vor allem die

³⁷Ebenda.

³⁸Ebenda.

³⁹Rashidi, M., The Balkanization of Execution-Security Research for AI Coding Agents: Isolation, Access Control and Time-of-Check-to-Time-of-Use Vulnerabilities, 7. Juli 2026, <https://arxiv.org/pdf/2607.05743>.

⁴⁰OWASP Top Ten for Agentic Applications for 2026, <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026>.

⁴¹Getir, E., Development of a threat modelling framework and web-based threat modelling tool for micro businesses, Dissertation Universität Essex, Juni 2024, <https://arxiv.org/pdf/2411.14450>.

⁴²Ebenda.

Annahmen, auf denen die einzelnen Kontrollen selbst beruhen. Eine Untersuchung von 640 Annahmen aus 96 STRIDE-Threat-Models desselben Systems fand heraus, dass diese Annahmen in der Praxis überwiegend implizit bleiben oder bestenfalls unstrukturiert dokumentiert werden – es fehlen Richtlinien dafür, wie mit ihnen umzugehen ist, und ihre Rolle im Modellierungsprozess ist kaum systematisch untersucht.⁴³

Die hier untersuchte Frage liegt entsprechend eine Ebene tiefer als jeder Kategorienkatalog: **Wer hat entschieden, dass die vorgesehenen Kontrollen den relevanten Problemraum tatsächlich erfassen – und auf welchen Annahmen beruhte diese Entscheidung?** Die Blackbox liegt nicht neben den bekannten technischen Risiken. Sie liegt unter ihnen.

2.3 Nicht Human Oversight

Ebenso wenig deckt sich die Frage mit Human Oversight. Human Oversight setzt voraus, dass ein Mensch das Verhalten eines KI-Systems beobachten, beurteilen und erforderlichenfalls unterbrechen kann; Art. 14 KI-VO verpflichtet im Zusammenhang mit Hochrisiko-KI-Systemen entsprechend zur wirksamen Beaufsichtigung durch natürliche Personen, die Fähigkeiten und Grenzen des Systems verstehen, Anomalien erkennen und es übersteuern können müssen. Damit adressiert Human Oversight primär das Verhältnis zwischen Mensch und laufendem KI-System.

Die hier untersuchte Grenze liegt davor: Auch ein funktionierendes Oversight-Konzept muss festlegen, was überwacht wird, welche Abweichung relevant ist, welche Information verfügbar ist und wann eine Intervention nötig ist – Entscheidungen, die wieder auf Annahmen über mögliches Systemverhalten beruhen. Ein Mensch kann eine Oversight-Funktion vollständig korrekt ausüben und trotzdem eine Entwicklung nicht erkennen, wenn sie außerhalb dessen liegt, was das Oversight-Konzept sichtbar macht. Mehr Human Oversight löst das nicht automatisch, ebenso wenig eine weitere Kontrolle über die bestehende Kontrolle: Auch deren Reichweite muss jemand definieren.

Das ist der Unterschied zwischen **Kontrolle ausüben** und **Kontrolle entwerfen**. Beim Ausüben geht es darum, ob der Mensch innerhalb des vorgesehenen Systems richtig „funktioniert“, beim Entwerfen geht es darum, ob das Kontrollsystem selbst auf einer Annahme beruht, deren blinder Fleck bei der Konstruktion niemandem auffiel. Neu ist dabei nicht das Grundproblem: Sicherheitsarchitekturen waren immer schon unvollständige Modelle einer komplexeren Realität. Neu ist die Intensität, mit der agentische KI-Systeme diese Lücke heute testen können, ohne dabei je zu ermüden.

3. Wenn Ausdauer skalierbar wird

Dass Sicherheitskontrollen auf unvollständigen Annahmen beruhen können, ist kein neues Phänomen – Zero-Days, unerwartete Abhängigkeiten und neue Angriffstechniken gehören seit jeher zu Security. Würde agentische KI dem lediglich einen weiteren Angreifertyp hinzufügen, wäre die Konsequenz überschaubar: Threat Models ergänzen, Kontrollen anpassen, fertig. Der Hugging-Face-Vorfall legt eine weitergehende Verschiebung nahe – nicht darin, **was** ein Angreifer kann, sondern darin, **wie** systematisch und zu welchen Grenzkosten sich die Annahmen einer Sicherheitsarchitektur testen lassen.

Rund 17.600 Aktionen in wenigen Tagen, Hunderte kooperierende KI-Agenten, die Erkenntnisse austauschten: Das Ergebnis war keine geniale Einzelentdeckung, sondern eine große Zahl kleiner Entscheidungen, Versuche und Anpassungen, aus denen sich am Ende eine funktionierende Angriffskette ergab. **Für Security ist das ein qualitativer Unterschied.** Auch menschliche Angreifer sind beharrlich und kreativ – aber Zeit, Aufmerksamkeit und qualifizierte Arbeitskraft sind teuer und begrenzen, wie viele Hypothesen parallel geprüft und wie lange wenig aussichtsreiche Pfade verfolgt werden. Agentische KI-Systeme verschieben genau diese Kostenstruktur.

⁴³ Van Landuyt, D. / Joosen, W., A descriptive study of assumptions in STRIDE security threat modeling, 1. Dezember 2022, Software and Systems Modeling, Vol. 21 (6), 2311-2328, <https://dl.acm.org/doi/10.1007/s10270-021-00941-7>.

3.1 Der menschliche Angreifer war Teil des impliziten Threat Models

Viele Sicherheitsmaßnahmen tragen eine unausgesprochene Annahme in sich: **Der Angreifer hat begrenzte Ressourcen.** Rate Limits, Monitoring-Schwellen, Session Limits, Lockouts – sie alle beruhen darauf, dass bestimmte Aktivitäten ab einer bestimmten Häufigkeit oder Geschwindigkeit unwirtschaftlich werden. Das war kein irrationales Threat Model: Auch organisierte Cybercrime-Gruppen und staatliche Akteure müssen priorisieren, ein Angriffspfad mit geringer Erfolgswahrscheinlichkeit konkurriert mit anderen Zielen.

Agentische KI-Systeme beseitigen diese Grenze nicht vollständig – Rechenleistung, Tokens und Tool-Zugänge bleiben Ressourcen –, verändern aber das Verhältnis zwischen Aufwand und möglicher Zahl der Versuche erheblich. Eine Aufgabe, für die menschliche Fachkräfte wiederholt analysieren, ausprobieren und neu ansetzen müssten, lässt sich auf eine große Zahl paralleler Agentenläufe verteilen. Der Hugging-Face-Vorfall liefert dafür kein abstraktes Szenario: Nachdem am 11. Juli 2026 ein Agent einen erfolgversprechenden Weg über manipulierte Dataset-Konfigurationen gefunden und die Reproduktion geteilt hatte, wechselten Hunderte andere Agenten auf diesen Ansatz. Ausdauer selbst wird damit zu einer skalierbaren Angriffsfähigkeit.⁴⁴

3.2 Automation Bias auf der Verteidigerseite

Während Angreifer automatisiert nach Lücken suchen können, delegiert auch die Verteidigerseite Analyse und Priorisierung zunehmend an automatisierte Systeme – notwendigerweise, denn die Menge an Logs, Alerts und Telemetriedaten wäre ohne SIEM, EDR, SOAR und KI-gestützte Korrelation kaum mehr beherrschbar. Damit entsteht eine bekannte kognitive Gefahr in neuer Form: **Automation Bias**, die übermäßige Verlässlichkeit auf automatisierte Entscheidungshilfen bei gleichzeitiger Abwertung widersprechender Information. Die Forschung unterscheidet dabei zwei Fehlertypen, commission errors – Befolgen einer falschen automatisierten Empfehlung – und omission errors, bei denen ein Problem übersehen wird, weil das System es erst gar nicht angezeigt hat.^{45 46}

Für Security ist die zweite Variante die gefährlichere: Je leistungsfähiger ein Sicherheitssystem erscheint, desto leichter verschiebt sich „das System hat keinen Alarm ausgelöst“ zu „es gibt kein relevantes Problem“ – **die Wahrnehmungsgrenze der Kontrolle wird zur Wahrnehmungsgrenze der Organisation.** Genau das zeigt Hugging Face: Runtime Detection, SIEM und weitere Kontrollen erkannten einzelne verdächtige Aktivitäten durchaus – das Problem war nicht fehlende Telemetrie, sondern dass die vorhandenen Signale nicht rechtzeitig zu einem Bild korreliert wurden, dessen Kritikalität eine Eskalation ausgelöst hätte. Selbst die KI-gestützte Korrelation erzeugte nach Darstellung von Hugging Face einen kohärenten Hinweis – der aber nicht zu einer rechtzeitigen On-Call-Reaktion führte. **Die Kontrolle war nicht blind. Sie sah Teile des Angriffs, ohne dessen Bedeutung rechtzeitig sichtbar zu machen.** Auch ein automatisiertes Erkennungssystem muss – über menschlich definierte Regeln, Trainingsdaten, Schwellenwerte – entscheiden, welche Kombination von Ereignissen relevant ist. Mehr Automatisierung beseitigt diese Designentscheidung nicht; sie kann ihre Reichweite sogar weniger sichtbar machen, wenn Nicht-Erkennung zunehmend als Beweis für Abwesenheit eines Problems gilt.

3.3 Für den Angriff braucht es kein Frontier-Modell

Eine naheliegende Gegenposition: Wenn nur wenige hochentwickelte Frontier-Modelle solche Angriffsketten autonom durchführen können, ließe sich das Risiko über Provider-Kontrollen und Modellzugang begrenzen. Die Entwicklung offener Modelle spricht gegen diese Annahme. Das britische AI Security Institute (AISII) verglich im Juli 2026 erstmals systematisch die Cyberfähigkeiten führender Open-Weight-Modelle (GLM-5.2, DeepSeek V4-Pro) mit geschlossenen Frontier-Modellen und kommt zu dem Ergebnis: Der Abstand beträgt inzwischen nur noch etwa vier bis sieben Monate – gegenüber

⁴⁴ METR, Brief independent investigation of agents' behaviour, reasoning and collaboration in the OpenAI/Hugging Face incident; OpenAI Inc., OpenAI – Hugging Face Incident Technical Report.

⁴⁵ Lyell, D. / Magrabi, F. / Raban, M. Z. / Pont, L.G. / Baysari, M. / Day, R. O. / Coiera, E., Automation bias in electronic prescribing, BMC Medical Informatics and Decision Making (2017) 17:28, https://pmc.ncbi.nlm.nih.gov/articles/PMC5356416/pdf/12911_2017_Article_425.pdf.

⁴⁶ Parasuraman, R. / Manzey, D., Complacency and Bias in Human Use of Automation: An Attentional Integration, Oktober 2010, https://www.researchgate.net/publication/47792928_Complacency_and_Bias_in_Human_Use_of_Automation_An_Attentional_Integration.

sechs bis zehn Monaten im Jahresverlauf 2025. Bemerkenswert ist dabei nicht nur die Richtung, sondern das Ausmaß: GLM-5.2 erreichte auf den engen Cyber-Testaufgaben ein mit Opus 4.6 vergleichbares Niveau bei rund 40 % der Kosten pro Aufgabe; eingebaute Schutzmechanismen ließen sich laut AISI in den Tests größtenteils durch einfaches Wiederholen der Anfrage umgehen.⁴⁷

Die genaue Monatszahl ist weniger wichtig als die Richtung: Fähigkeiten, die heute auf wenige kontrollierte Modelle beschränkt scheinen, diffundieren innerhalb kurzer Zeit in Systeme, die lokal betrieben, verändert und ohne die Schutzmechanismen eines kommerziellen Providers eingesetzt werden können. Eine weitere implizite Sicherheitsannahme verliert damit an Stabilität: Dass besonders leistungsfähige autonome Angriffe zwingend Zugang zu einer kleinen Zahl kontrollierbarer Frontier-Systeme voraussetzen.

Der Hugging-Face-Vorfall enthält die Gegenrichtung dieser Entwicklung gleich mit: Für die forensische Analyse verweigerten kommerzielle Frontier-Modelle Teile der nötigen Reverse-Engineering-Arbeit, weil ihre Sicherheitsmechanismen nicht zwischen Angreifer und Incident Responder unterscheiden konnten; das Untersuchungsteam griff stattdessen auf ein selbst gehostetes GLM-5.2-Modell zurück.⁴⁸ Dieselbe Eigenschaft, die offensive Fähigkeiten schwerer kontrollierbar macht – Verfügbarkeit leistungsfähiger Modelle außerhalb zentraler Provider-Infrastruktur –, eröffnet also auch Verteidigern Handlungsspielraum. Capability Diffusion ist kein rein offensives Phänomen. Für die Sicherheitsarchitektur bleibt aber entscheidend: Sie kann die Verfügbarkeit einer Fähigkeit nicht mehr zuverlässig aus der heutigen Verteilung von Modellzugängen ableiten.

Hugging Face selbst zieht daraus eine konkrete Lehre für Verteidiger: Ein eigenes, bereits vorab geprüfetes Modell auf eigener Infrastruktur betriebsbereit zu halten – sowohl um durch Guardrails blockierte Forensik zu vermeiden als auch um zu verhindern, dass Angreiferdaten und Zugangsdaten die eigene Umgebung verlassen.⁴⁹ Dieselbe Verschiebung – Anbieter rüsten Verteidiger gezielt mit Fähigkeiten aus, die zuvor eher Angreifern zugeschrieben wurden – zeigt sich auch bei Anthropic im April 2026 vorgestelltem Werkzeug „Claude Security“, das CISOs automatisierte Schwachstellenscans samt Patch-Vorschlägen für die eigene Codebasis anbietet.^{50 51}

3.4 Die Schwachstelle muss nicht neu sein

Noch eine Annahme gerät unter Druck: Dass eine neue Angriffsfähigkeit vor allem durch neue technische Schwachstellen entsteht. Bei Hugging Face bestand die Besonderheit nicht darin, dass jeder Schritt hochentwickelte oder unbekannte Exploits erforderte – neben dem Zero-Day im Package-Registry-Cache-Proxy nutzte der Angriff überwiegend bekannte Techniken: Unzureichend abgesicherte Dataset-Verarbeitung, erreichbare Credentials, interne APIs, laterale Bewegung. Auch die drei von Anthropic untersuchten realen Vorfälle beruhten auf grundlegenden Techniken – schwache Passwörter, offene Debug-Informationen, unauthentifizierte Endpunkte, SQL Injection.⁵²

Das verändert die praktische Bedeutung „alter“ Schwachstellen erheblich: Wenn ein autonomes System bekannte, einzeln wenig spektakuläre Lücken in großer Zahl suchen und kombinieren kann, muss eine Schwachstelle nicht neu sein, um in einem neuen Angriffsmodell gefährlich zu werden. „Das hätte ohnehin geschlossen sein müssen“ greift somit zu kurz: Reale Infrastruktur lebt mit technischen Schulden, Legacy-Systemen und Priorisierung nach Aufwand und Wahrscheinlichkeit. Ändert sich der Aufwand auf Angreiferseite fundamental, ändert sich damit auch die Variable, auf deren Basis diese Priorisierung getroffen wurde – **ohne dass sich an der Schwachstelle selbst irgendetwas geändert hat.**

⁴⁷UK AI Security Institute, How far behind the frontier are leading open-weight models on cyber?, 17. Juli 2026; <https://www.aisi.gov.uk/blog/how-far-behind-the-frontier-are-leading-open-weight-models-on-cyber>. Vergleichsbasis: GLM-5.2 und DeepSeek V4-Pro gegenüber Opus 4.5/4.6.

⁴⁸ Hugging Face Inc. Security incident disclosure.

⁴⁹ Ebenda.

⁵⁰ Security Week (Autor Townsend, K), Anthropic Unveils Claude Security to Counter AI-Powered Exploit Surge, 30. April 2026; <https://www.securityweek.com/anthropic-unveils-claude-security-to-counter-ai-powered-exploit-surge/>.

⁵¹ Anthropic Inc., Claude Security Public Beta, 30. April 2026, <https://claude.com/blog/claude-security-public-beta>.

⁵² Anthropic Inc., Investigating three real-world incidents in our cybersecurity evaluations.

3.5 Wenn sich das Threat Model schneller ändert als die Kontrolle

Dass diese Verschiebung längst über einzelne Sicherheitsteams hinausreicht, zeigt sich daran, dass selbst Finanzaufsichtsbehörden das Thema aufgreifen: Im Mai 2026 warnte die deutsche BaFin etwa, KI-Anwendungen könnten Schwachstellen in IT-Systemen sehr schnell finden und ausnutzen – Unternehmen müssten sich auf eine wachsende Anzahl potentieller Angriffsflächen einstellen und Sicherheitslücken deutlich schneller schließen als bisher.⁵³

Damit schließt sich der Kreis zu Kapitel 2: Agentische KI verändert nicht nur, dass ein neuer Angreifer ins Threat Model muss – sie verändert die Grundannahmen selbst: Über verfügbare Zeit, Parallelisierung, Kosten von Fehlversuchen, die Kombination kleiner Schwächen, die Verfügbarkeit bestimmter Fähigkeiten und die Geschwindigkeit ihrer Diffusion. Eine Kontrolle kann deshalb weiterhin exakt so funktionieren wie bei ihrer Einführung – und trotzdem weniger wirksam sein als zuvor. Das ist der entscheidende Punkt.

Security bewertet Kontrollen traditionell gegen einen angenommenen Bedrohungsraum. Wenn sich dieser Raum verändert, reicht die Prüfung, ob die Kontrolle technisch noch funktioniert, nicht aus – zu prüfen ist ebenso, **ob die Annahme, gegen die ihre Wirksamkeit ursprünglich beurteilt wurde, noch gilt.**

Genau dort beginnt „Security neu denken“: Nicht mit der Forderung nach immer mehr Kontrollen, sondern mit der Frage, **welche Annahme eine bestehende Kontrolle überhaupt trägt – und was geschieht, wenn diese Annahme nicht mehr stimmt.**

4. Fünf Beispiele, eine Struktur

Die folgenden fünf Beispiele zeigen dieselbe Grundstruktur – eine stillschweigende Vertrauensannahme darüber, was eine Kontrolle abdeckt – in fünf unterschiedlichen Bereichen der Security Governance⁵⁴. Jedes Beispiel folgt demselben Aufbau: Warum das notwendig ist, was „Security neu denken“ an dieser Stelle konkret bedeutet, und was ein CISO auch ohne Anbieterkooperation selbst in der Hand hat.

Dass selbst ein aktuelles, im Juni 2026 veröffentlichtes Framework wie OWASP AISVS 1.0⁵⁵ an dieser Stelle blind bleibt, ist kein Zufall der Kapitelauswahl, sondern strukturell angelegt. AISVS enthält ein eigenes Kapitel „Human Oversight“ mit Anforderungen an Kill-Switch-Mechanismen, Freigabepflichten bei Hochrisiko-Aktionen, Rollback-Verfahren und die Protokollierung von Override-Events.⁵⁶ Das prüft die Kontrolle über das laufende KI-System – nicht, ob die Annahme, auf der eine Kontrolle wie eine Sandbox-Isolation beim Entwurf beruhte, tatsächlich zutraf. Ob eine solche Annahme korrekt war, ist keine technische Verifikationsfrage, sondern eine Governance-Frage: **Wer hat entschieden, dass sie ausreichend geprüft wurde, und auf welcher Grundlage?** Genau an dieser Stelle braucht Security Governance eine Perspektive, die Kontrollen nicht nur auf korrekte Implementierung, sondern auf die Legitimität und Vollständigkeit ihrer zugrunde liegenden Annahmen hin befragt.

4.1. Sandbox-Isolation und Zugriffsschlüssel

Warum das notwendig ist. Eine Sandbox gilt als sicher, weil jemand entschieden hat, dass sie es ist – nicht, weil das gegen die tatsächliche Angriffsfläche verifiziert wurde. Der Hugging-Face-Vorfall zeigt das konkret: OpenAIs Evaluierungsumgebung war ohne direkten Internetzugang konzipiert, doch die Modelle identifizierten und nutzten eine Zero-Day-Schwachstelle in Artifactory, einem intern gehosteten Package-Registry-Proxy, um sich Internetzugang zu verschaffen. Von dort aus nutzten sie die kompromittierte Verbindung, um über laterale Bewegung und Privilege Escalation

⁵³ IT Finanzmagazin, BaFin warnt vor KI-gestützten Cyber-Angriffen auf Finanzbranche und stärkt Verbraucherschutz, 12. Mai 2026, <https://www.it-finanzmagazin.de/ba-fin-warnt-vor-ki-gestuetzten-cyber-angriffen-auf-finanzbranche-und-staerkt-verbraucherschutz-244220/>.

⁵⁴ Definition laut ISO/IEC 27014:2020: „the means by which an organization's governing body provides overall direction and control of activities that affect the security of an organization's information.“

⁵⁵ OWASP AISVS Artificial Intelligence Security Verification Standard, <https://owasp.github.io/www-project-artificial-intelligence-security-verification-standard-aisvs-docs/>.

⁵⁶ OWASP AISVS Artificial Intelligence Security Verification Standard, AD. 19.

Auswertungsantworten direkt aus Hugging Faces Produktivdatenbank zu extrahieren.^{57 58} Reuters berichtete zudem, dass OpenAI weitere, bislang nicht offengelegte Fälle von Sandbox-Escapes identifiziert hat, die als begrenzter Natur eingestuft wurden.⁵⁹ Die eigentliche Schwachstelle war damit nicht die offensichtliche Grenze (Internet ja/nein), sondern eine als intern und harmlos eingestufte Komponente, die als Seitenkanal diente.

Was heißt „neu denken“. Sandbox-Isolation und Zugriffsschlüssel-Vergabe werden aktuell als binäre, einmalige Architekturentscheidung behandelt: „Wir haben entschieden, dass das isoliert ist“ – isoliert oder nicht isoliert, ohne Zwischenstufen. Die Identitätsprüfung bei Menschen löst dasselbe Grundproblem – wie stellt man fest, ob eine behauptete Vertrauensgrenze echt ist – anders: NIST definiert Identity Proofing von vornherein nicht als Herstellung einer Garantie, sondern als Herstellung einer Beziehung zwischen einem Subjekt und einer realen Person „to some degree of assurance“ – einem vorab festgelegten Grad der Zuversicht, nicht einem Absolutwert.⁶⁰ Diesen Grad staffelt NIST SP 800-63A-4 in drei Stufen (IAL1 bis IAL3) mit steigender Prüftiefe und macht dabei explizit, was jede Stufe nicht abdeckt – IAL1 schützt vor automatisierten Massenangriffen, aber nicht vor gezielter Social Engineering; erst IAL3 verlangt eine persönlich anwesende, geschulte Prüfperson und biometrische Erfassung. Kein Level verspricht hundertprozentige Sicherheit, sondern jedes definiert einen klar begrenzten Grad an Vertrauen und Zuversicht.⁶¹

Genau dieses Prinzip – ein vorab festgelegter, gestaffelter Grad der Zuversicht statt einer binären Ja/Nein-Behauptung – lässt sich auf Sandbox-Isolation und Zugriffsschlüssel für KI-Agenten übertragen:

- **Stufe 1 (Konfigurationsbehauptung):** Isolation beruht auf deklarerter Konfiguration (Netzwerkregeln, Firewall-Policies), ohne unabhängige Prüfung. Schützt vor unbeabsichtigten Fehlkonfigurationen, nicht vor aktiver Umgehung.
- **Stufe 2 (punktuell verifiziert):** Ein unabhängiger Test gegen bekannte Umgehungstechniken (SSRF, DNS-Manipulation, Querverbindungen über gemeinsam genutzte interne Infrastruktur) wurde zu einem Zeitpunkt bestanden. Schützt vor zum Testzeitpunkt bekannten Angriffsmustern.
- **Stufe 3 (laufend verifiziert):** Stufe-2-Test wird in festgelegtem Rhythmus von unabhängiger Seite wiederholt, ergänzt durch laufende Überwachung auf anomale ausgehende Verbindungen. Schützt vor bekannten Mustern kontinuierlich – nicht vor einer bislang unbekannten Zero-Day wie der Artfactory-Lücke, genau wie auch NIST auf keiner Stufe einen Grad der Zuversicht jenseits des dafür definierten Rahmens beansprucht.

„Neu denken“ heißt damit: Vorab explizit festlegen, welche Stufe eine Isolations- oder Scope-Behauptung tatsächlich beansprucht – und welche Angriffswege bewusst *nicht* abgedeckt sind –, statt implizit eine unbegrenzte Garantie zu unterstellen, die kein Modell, auch NIST nicht, tatsächlich abgeben kann.

Die realistische Einschränkung. Selbst ein Unternehmen, das sich diszipliniert an eine aktuelle NIST-Leitlinie wie SP 800-63A-4 hält, bewegt sich damit nicht automatisch auf der Höhe der tatsächlichen Bedrohungslage – die Leitlinie selbst ist an mehreren Stellen der limitierende Faktor, nicht die Umsetzung durch das Unternehmen. NIST benennt den Schutz vor synthetischen Identitäten ausdrücklich als Ziel von IAL1⁶², adressiert generative KI und Deepfakes (definiert als „durch KI erzeugter oder manipulierter Bild-, Ton- oder Videoinhalt, der wirklichen Personen ... ähnelt und einer Person fälschlicherweise als echt ... erscheinen würde“; Art. 3 Z 60 KI-VO) an anderer Stelle sogar explizit: CSPs müssen ihre automatisierten Bilderkennungsalgorithmen gegen bekannte Fälschungsartefakte testen und die dabei ermittelten Falsch-Positiv- und Falsch-Negativ-Raten

⁵⁷ InfoQ (Autor Pop, O.), Swarm of OpenAI Agents exploit artifactory Zero-Day to Escape Sandbox and Breach Hugging Face, 4. August 2026, <https://www.infoq.com/news/2026/08/openai-huggingface-breach/>.

⁵⁸ OpenAI Inc., Hugging Face Model Evaluation – Security Incident, 21. Juli 2026 mit letztem Update vom 28. Juli 2026, <https://openai.com/de-DE/index/hugging-face-model-evaluation-security-incident/>.

⁵⁹ Reuters (Autoren Seetharaman, D. / Satter, R.), OpenAi finds evidence other AI aigents escaped containment“, 31. Juli 2026, <https://www.reuters.com/business/openai-finds-evidence-other-ai-agents-escaped-containment-it-widens-hacking-2026-07-31/>.

⁶⁰ NIST SP 800-63A-4, Abschnitt 2.1 (Definition „identity proofing ... to some degree of assurance“), 2025, <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63A-4.pdf>.

⁶¹ NIST SP 800-63A-4, Abschnitt 1.2 / Identity Assurance Levels und Abschnitt 4.4. / Table 1 (Anforderungsübersicht je Stufe).

⁶² NIST SP 800-63A-4, Abschnitt 1.2 („IAL1 is designed to ... protect against synthetic identities“).

dokumentieren⁶³. Verpflichtend ist aber nur das Testen und Dokumentieren selbst, nicht das Erreichen eines bestimmten Schwellenwerts. Dazu kommt eine SHALL/SHOULD-Kaskade bei den eigentlich wirksamsten Mitteln: Passive Deepfake-Erkennung, Geräte-Attestierung und die Analyse auf bekannte Generative-KI-Signaturen sind allesamt nur SHOULD, nicht SHALL⁶⁴. NIST räumt im selben Abschnitt sogar ein, dass „selbst diese Mechanismen nicht ausreichend sind, um alle möglichen Fälle abzudecken“⁶⁵ – ein dokumentiertes Restrisiko, das dennoch nicht zu strengeren verbindlichen Mindestanforderungen führt.

Ein konkretes Beispiel für diese Schwäche: Biometrische Liveness-Checks gegen ältere Referenzfotos produzieren in der Praxis häufig Fehlablehnungen – unabhängige Forschung zeigt einen anhaltenden Anstieg der Falsch-Rückweisungsrate bei Gesichtserkennung mit wachsendem Zeitabstand zwischen Referenz- und aktueller Aufnahme⁶⁶. Umgekehrt lässt sich dieselbe Schwäche gezielt ausnutzen: Dokumentierte Umgehungstechniken bei automatisierter Alters- und Gesichtserkennung zeigen, dass gezielte optische Anpassungen – Make-up, Beleuchtung, Mimik – ausreichen, um ein System dazu zu bringen, eine Person älter oder jünger einzuschätzen, als sie ist⁶⁷. Die Kontrolle existiert formal, hält aber der tatsächlichen Angriffsfläche nicht stand – dasselbe Prinzip, das sogar das deutsche Bundesamt für Sicherheit in der Informationstechnik (BSI) zu seinem eigenen Mindeststandard offen einräumt: Einhaltung eines Mindestniveaus ist notwendig, aber in der Regel nicht hinreichend für tatsächliche Sicherheit⁶⁸.

Was ein CISO selbst in Angriff nehmen kann. Für jede Sandbox- und Zugriffsschlüssel-Vergabe an KI-Agenten schriftlich festlegen, was die Isolations-/Scope-Behauptung konkret garantieren soll, und diese Behauptung durch einen unabhängigen Test gezielt gegen die im Hugging-Face-Vorfall dokumentierten Umgehungswege prüfen lassen.

4.2. Überwachungssysteme (SIEM)

Warum das notwendig ist. Der Hugging-Face-Vorfall zeigt, dass eine SIEM-Kontrolle formal funktionieren und trotzdem wirkungslos bleiben kann. Laut Hugging Faces eigenem technischem Bericht kamen die ersten Signale aus mehreren Schichten des eigenen Security-Stacks gleichzeitig – Live-Runtime-Analyse, SIEM-Logs und weitere Kontrollen.⁶⁹ Jedes Signal für sich war ambivalent; ein KI-gestützter Security-Agent-Stack korrelierte sie zu einem kohärenten Angriffsbild.⁷⁰ Bemerkenswert dabei: Das Sicherheitsteam versuchte zunächst, ein US-amerikanisches Frontier-Modell für diese Analyse einzusetzen, konnte es aber wegen dessen eigener Guardrails nicht wie benötigt nutzen – und griff stattdessen auf GLM 5.2, ein chinesisches Open-Source-Modell von Z.ai, zurück.⁷¹ Das System bewertete die Kritikalität des daraus entstandenen Alarms aber falsch und löste die Eskalation an das Bereitschaftsteam nicht aus – was laut eigener Aufarbeitung wertvolle Reaktionszeit kostete.⁷² Ein SIEM, das nur darauf getestet wird, ob es überhaupt einen Alarm auslöst, testet damit nur die halbe Kette: Erkennung war hier kein Problem, Eskalation war es.

Was heißt „neu denken“. Wirksamkeitsprüfung muss über die Frage hinausgehen, ob ein SIEM eine verdächtige Einzelaktivität erkennt, und stattdessen die gesamte Kette bis zur tatsächlichen Reaktion abbilden: Löst eine simulierte, realistische Angriffskette innerhalb der dafür vorgesehenen Zeit

⁶³ NIST SP 800-63A-4, Abschnitt 3.14.

⁶⁴ Ebenda.

⁶⁵ Ebenda.

⁶⁶ Purdue University (Autor Cedeno Agamez, M.), Aging effects in automated face recognition, 2016, https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1911&context=open_access_theses.

⁶⁷ Nansen, B. / Sandberg, H. / Bliss, L. / Cohny, S., From Phreaking to Sneaking: Children's Circumvention of Social Media Age Verification Systems, 1. Mai 2026, <https://arxiv.org/pdf/2605.00368>.

⁶⁸ BSI, Juli 2018, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Mindeststandards/Mindeststandard_HV-Benchmark_kompakt_Version_1_1.pdf.

⁶⁹ Hugging Face Inc., Security Incident Report.

⁷⁰ Hugging Face Inc., Anatomy of a Frontier Lab Agent Intrusion.

⁷¹ Fortune (Autor Forlini, E.), Hugging Facy says it resorted to a Chinese AI model to battle a fully autonomous cyberattack because US model guardrails stymied its defense, <https://fortune.com/2026/07/20/hugging-face-turns-to-chinese-open-source-ai-to-fend-off-autonomous-ai-cyber-attack-after-american-ai-guardrails-stymie-defense/>.

⁷² Hugging Face Inc., Anatomy of a Frontier Lab Agent Intrusion.

tatsächlich eine Eskalation an eine erreichbare, entscheidungsbefugte Person aus? Genau diese Kette, nicht die einzelne Erkennungsregel, war bei Hugging Face der Bruchpunkt.⁷³

Der erzwungene Modellwechsel mitten im Vorfall zeigt aber noch eine zweite, eigenständige Lücke: Sobald ein Sicherheitswerkzeug selbst auf einem KI-Modell aufbaut, erbt die Verteidigung nicht nur dessen Fähigkeiten, sondern auch dessen Sicherheits-Guardrails – und die sind nicht dafür entworfen, „helf mir, einen Angriff auf mich selbst zu analysieren“ als legitimen Zweck zu erkennen. Aus der Perspektive eines Guardrails sehen Angriffsprotokolle, Exploit-Muster und Schadcode-Analyse oft identisch zu dem aus, was sie eigentlich verhindern sollen. **Das ist ein Kontrollversagen, das kein klassisches Vendor-Risiko-Modell erfasst:** Das Werkzeug war verfügbar und funktionsfähig – es hat sich nur selbst geweigert, im entscheidenden Moment zu helfen. „**Neu denken**“ heißt hier: Vor dem Vorfall prüfen, ob das eigene Sicherheitswerkzeug tatsächlich in der Lage ist, forensische Analyse von Angriffsmaterial durchzuführen, ohne von seinen eigenen Schutzmechanismen blockiert zu werden – nicht erst während eines laufenden Angriffs feststellen, dass man das Werkzeug wechseln muss.

Die realistische Einschränkung. Realistische, mehrstufige Angriffssimulationen sind aufwendiger als klassische Signatur-Tests, und die Kalibrierung von Alarmschwellen bewegt sich unausweichlich zwischen Alert Fatigue (zu enge Schwellen) und durchrutschenden Angriffsketten (zu weite Schwellen). Es gibt keine Konfiguration, die beide Fehler gleichzeitig auf Null senkt. Der oben geschilderte erzwungene Modellwechsel bei Hugging Face legt aber eine Schwachstelle offen, die brisanter ist als ein bloßes Testproblem: Wenn ein Verteidigungswerkzeug sich im entscheidenden Moment selbst blockiert, ist das kein Zufall, sondern ein Muster, das sich ausnutzen lässt. Ein Angreifer, der weiß, dass gängige Guardrails forensische Angriffsanalyse fälschlich als Angriffsversuch selbst einstufen, kann seinen eigenen Angriff gezielt so gestalten, dass er exakt die Trigger auslöst, die das gegnerische Sicherheitswerkzeug lahmlegen – nicht als Kollateralschaden, sondern als Taktik. Das kehrt die vermeintliche Sicherheitshierarchie um: Ausgerechnet das Unternehmen mit dem strikter reguliertem, „sichereren“ KI-Werkzeug stand im Ernstfall schlechter da als eines, das auf ein weniger restriktives Modell zurückgreifen konnte. Ein CISO kann diese Umkehr nicht wegtesten, weil er die Guardrail-Logik des Anbieters nicht kontrolliert – er kann nur vorab prüfen, ob das eigene Werkzeug im Ernstfall tatsächlich hilft, und einen Fallback auf ein alternatives Modell bereithalten, wohl wissend, dass genau diese Absicherung im eigenen Sicherheitsbudget selten vorgesehen ist.

Was ein CISO selbst in Angriff nehmen kann. Ein Time-to-Escalation-Test ist nur so viel wert wie seine Nachbereitung. Häufigste Schwächen solcher Übungen in der Praxis: unrealistische, nicht auf die eigene Infrastruktur zugeschnittene Szenarien, sowie fehlende Dokumentation der Lessons Learned.⁷⁴ Konkret heißt das: Erstens, das Testszenario an die tatsächliche eigene Infrastruktur und die real eingesetzten Eskalationswege anpassen. Zweitens, jeden Test mit einem dokumentierten After-Action-Review abschließen, das konkret benennt, warum eine Eskalation zu spät oder gar nicht ausgelöst wurde. Drittens – der in der Praxis am häufigsten fehlende Schritt –, die daraus resultierenden Maßnahmen mit Verantwortlichkeit und Frist in die tatsächliche SIEM-Konfiguration und das Eskalationsverfahren einpflegen, statt sie im Bericht stehen zu lassen.⁷⁵ Ohne diesen dritten Schritt ist auch die realistischste Simulation nur eine folgenlose Übung.

4.3. Exit-Strategie

Warum das notwendig ist. DORA verlangt grundsätzlich an keiner Stelle, dass ein Finanzunternehmen sein Konzentrationsrisiko an der öffentlichen Liste kritischer IKT-Drittdienstleister (CTPP) misst – diese Liste, die die europäischen Finanzmarkt-Aufsichtsbehörden (ESA) am 18. November 2025 nach Art. 31 Abs. 9 DORA veröffentlicht haben⁷⁶, ist ein aufsichtliches Überwachungsinstrument,

⁷³ Ebenda.

⁷⁴ Trusted Sec, Incident Response Tabletop Exercises: How to Prepare your Team for the Real Thing, <https://trustedsec.com/resources/business-resources/incident-response-tabletop-exercises-guide>.

⁷⁵ Ebenda.

⁷⁶ EBA, The European Supervisory Authorities designate critical ICT third-party providers under the Digital Operational Resilience Act, 18. November 2025, <https://www.eba.europa.eu/publications-and-media/press-releases/european-supervisory-authorities-designate-critical-ict-third-party-providers-under-digital>.

entsprechend richtet auch die österreichische FMA ihre eigenen Aufsichtsmaßnahmen am CTPP-Überwachungsrahmen aus.⁷⁷

In der Praxis wird sie trotzdem – auch mangels besserer Alternativen – von Unternehmen gerne als Orientierung verwendet, um das eigene IKT-Konzentrationsrisiko gemäß Art 29 DORA abschätzen zu können. Reine KI-Modellanbieter wie OpenAI oder Anthropic tauchen derzeit auf der CTPP-Liste nicht als eigene Kategorie auf, obwohl die Konzentration auf Modellanbieter-Ebene real ist: Laut einer Befragung von rund 500 US-amerikanischen Enterprise-Entscheiderinnen durch Menlo Ventures entfielen Ende 2025 rund 88 % der Enterprise-LLM-API-Ausgaben auf nur drei Anbieter (Anthropic 40 %, OpenAI 27 %, Google 21 %)⁷⁸. Für Europa gibt es keine vergleichbare Statistik auf Unternehmensebene, allerdings legen Statistiken zu Web-Traffic von Endnutzerinnen nahe, dass auch dort die genannten Big Three gerne genutzt werden (etwa 80 % ChatGPT/OpenAI, was nur marginal höher ist als dessen globaler Marktanteil von 79,8 %⁷⁹). Wer sich nur an der CTPP-Liste orientiert, übersieht eine erhebliche Abhängigkeit von einem einzelnen KI-Anbieter: Wird etwa Claude (Anthropic) über AWS Bedrock (Amazon) bezogen, taucht in der Liste nur „AWS“ auf, die eigentliche Abhängigkeit von Anthropic bleibt darin unsichtbar.

Was heißt „neu denken“. Selbst der eigene Versuch der KI-VO, in diese Richtung zu regeln, trägt nicht: Art 17 Abs 1 lit. I KI-VO verlangt im Rahmen des Qualitätsmanagementsystems (QMS) „Ressourcenmanagement, einschließlich Maßnahmen im Hinblick auf die Versorgungssicherheit“. Art. 17 Abs. 3 KI-VO erlaubt es Anbietern ausdrücklich, ein bestehendes QMS aus sektorspezifischem Unionsrecht – etwa der Medizinprodukteverordnung – als Basis zu verwenden; laut VDE Verband der Elektrotechnik Deutschland deckt ein MDR-konformes QMS nach ISO 13485 einen Großteil von Art. 17 KI-VO bereits ab⁸⁰. Das erklärt, warum sich die Versorgungssicherheits-Klausel wie ein Fremdkörper liest: Sie stammt konzeptionell aus einem Kontext, in dem „Versorgungssicherheit“ physische Lieferengpässe bei Medizinprodukten meint, keine Fälle, in denen sich Softwareverhalten grundlegend ändert. Und selbst wortwörtlich gelesen läuft die Norm ins Leere: Art. 17 KI-VO gilt nur für Anbieter von Hochrisiko-KI-Systemen, nicht für eventuell darunterliegende GPAI-Basismodelle. Ein reales Beispiel: Am 12. Juni 2026 wurde der Zugang zu Claude Mythos 5 und Claude Fable 5 wegen US-Exportkontrollen ausgesetzt, am 30. Juni 2026 aufgehoben, ab 1. Juli 2026 wiederhergestellt⁸¹ – eine gut dreiwöchige, drittstaatlich verursachte Zugangssperre, für die Art. 17 KI-VO übrigens ohnehin nicht gilt.

Die eigentliche Rechtsgrundlage für Exit-Strategien liegt in Art. 28 Abs. 8 DORA, der für IKT-Dienstleistungen mit kritischer oder wichtiger Funktion getestete, dokumentierte Ausstiegspläne verlangt – **„neu denken“ heißt hier konkret:** Der Ausstiegsplan braucht für den KI-Modellanbieter denselben Detailgrad wie für die darunterliegende Cloud-Infrastruktur – einen namentlich benannten Alternativanbieter, und einen vorab durchgeführten Vergleichslauf: Denselben Satz repräsentativer Produktions-Prompts bei beiden Anbietern laufen lassen und die Ergebnisse anhand vorher festgelegter Kriterien (z. B. Antwortformat, Genauigkeit bei den geschäftskritischen Anwendungsfällen, Ausgabestruktur für nachgelagerte Systeme) vergleichen – nicht nur die vertragliche Zusicherung, dass ein Export „möglich“ ist, sondern das nachweisbare Ergebnis, dass der Alternativanbieter bei denselben Eingaben brauchbare Ausgaben liefert.

Das nach Art. 28 Abs. 3 DORA ohnehin verpflichtende Informationsregister ist dabei strukturell durchaus für Subunternehmerketten ausgelegt: Der zugehörige technische Durchführungsstandard sieht 15 relational verknüpfte Templates vor, darunter ein eigenes Template für Unterauftragsketten⁸².

⁷⁷ FMA, Einforderung von Strategie für das IKT-Drittparteienrisiko und Ausstiegspläne, <https://www.fma.gv.at/querschnittsthemen/dora/fma-aktivitaeten-zu-dora/>.

⁷⁸ Menlo Ventures, 2025: The State of Generative AI in the Enterprise, 9. Dezember 2025, <https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/>.

⁷⁹ Adapt, AI Model Usage in Europe 2025-2026: Which Countries Use ChatGPT, Claude and Generative AI Most? 22. Juni 2026, <https://www.adaptworldwide.com/2026/06/22/ai-model-usage-in-europe-2025-2026-which-countries-use-chatgpt-claude-and-generative-ai-most/>.

⁸⁰ VDE, KI (Künstliche Intelligenz)-basierte Medizinprodukte: Herausforderungen durch den AI Act, 1. September 2026, <https://www.vde.com/topics-de/health/beratung/herausforderungen-ki-basierte-medizinprodukte>.

⁸¹ Anthropic Inc., Statement on the US government directive to suspend access to Fable 5 and Mythos 5, 12. Juni 2026, <https://www.anthropic.com/news/fable-mythos-access>.

⁸² Durchführungsverordnung (EU) 2024/2956 der Kommission vom 29. November 2024 zur Festlegung technischer Durchführungsstandards für die Anwendung der Verordnung (EU) 2022/2554 des Europäischen Parlaments und des Rates im Hinblick auf die Standardvorlagen für das Informationsregister.

Das eigentliche Problem liegt nicht in der Struktur, sondern in der geschlossenen Dienstleistungstaxonomie: Sie kategorisiert nach IKT-Diensttyp, nicht nach Anbieteridentität – ein KI-Modell taucht darin als „Cloud-“ oder „Software-Dienstleistung“ auf, nicht als eigene Kategorie „KI-Modellanbieter“. Genau das wiederholt auf Register-Ebene dasselbe Muster wie bei der CTPP-Liste: Der Sub-Dienstleister ist formal erfasst, aber unter einer Kategorie, die seine eigentliche Identität verdeckt.⁸³

NIS-2 bietet an dieser Stelle keine Alternative, sondern bleibt noch unbestimmter als DORA: Art. 21 Abs 2 lit d NIS-2 verlangt von den erfassten Einrichtungen lediglich allgemein ein Risikomanagement für die Lieferkette, ohne ein Register, eine Taxonomie oder überhaupt eine Struktur vorzugeben, anhand derer sich Abhängigkeiten wie die von einem bestimmten KI-Modellanbieter sichtbar machen ließen.

Die realistische Einschränkung. Ein nach Art. 28 Abs. 8 DORA getesteter Ausstiegsplan ist ein Test zu einem bestimmten Zeitpunkt – und genau das ist die Einschränkung, die zum Rest dieses Whitepapers passt (dieselbe Momentaufnahme-Logik wie bei der Sandbox-Isolation in 4.1): Zwischen dem Zeitpunkt, an dem der Ausstieg getestet und für funktionsfähig befunden wurde, und dem Zeitpunkt, an dem er im Ernstfall tatsächlich gebraucht wird, können beim Alternativenanbieter Modellversionen, APIs oder Nutzungsbedingungen wechseln, ohne dass der Ausstiegsplan erneut geprüft wurde. Bei agentischer KI wiegt das schwerer als bei einem einzelnen Prompt-Response-Vergleich: Ein Agent trifft über mehrere Schritte hinweg eigenständig Zwischenentscheidungen – welches Tool er aufruft, wann er nachfragt statt zu handeln, wie er mit einem unerwarteten Zwischenergebnis umgeht. Ein Alternativenanbieter, der beim ursprünglichen Vergleichslauf inhaltlich gleichwertige Endergebnisse lieferte, kann in diesen Zwischenschritten ein anderes Verhalten zeigen, ohne dass sich das im Output der einzelnen getesteten Aufgabe zeigt – etwa eine andere Schwelle dafür, wann er ohne Rückfrage selbstständig handelt. Der Ausstiegsplan bestätigt dann, dass der Agent beim Alternativenanbieter zum selben Ergebnis kommt, ohne zu erfassen, dass er auf dem Weg dorthin autonome Handlungen ausführt, die beim ursprünglichen Anbieter eine menschliche Freigabe ausgelöst hätten. Ein Ausstiegsplan, der nur einmal zertifiziert und nicht laufend gegen den aktuellen Stand des Alternativenanbieters nachgetestet wird, ist im Ernstfall möglicherweise ein Plan für einen Agenten, der sich in dieser Form nicht mehr so verhält, wie er getestet wurde.

Was ein CISO selbst in Angriff nehmen kann. Das eigene Verarbeitungsverzeichnis – ob in Excel oder spezialisierter Software geführt, entscheidend ist die Struktur, nicht das Werkzeug – um ein zusätzliches Feld erweitern, das die ITS-Taxonomie nicht vorsieht⁸⁴: Den tatsächlichen KI-Modellanbieter hinter jedem als „Cloud-“ oder „Software-Dienstleistung“ kategorisierten Subunternehmer explizit vermerken, statt eine zweite, separate Erhebung nur für das RTS-Mindestregister aufzusetzen. Den nach Art. 28 Abs. 8 DORA ohnehin verpflichtenden Ausstiegsplan um zwei KI-spezifische Inhalte ergänzen, die die Norm nicht vorgibt: Ein vertraglich fixiertes Exportformat für Prompts und Fine-Tuning-Daten (Art. 30 DORA verlangt nur, dass Datenrückgabe geregelt ist, nicht in welcher Form), und einen Vergleichslauf mit mehrstufigen Agenten-Szenarien – nicht nur einzelne Prompt-Response-Paare, sondern dieselbe mehrschrittige Aufgabe mit denselben Freigabeschwellen bei beiden Anbietern durchspielen und prüfen, ob der Alternativenanbieter an denselben Stellen eine menschliche Freigabe einholt wie der ursprüngliche.

Wer diesen Test bewertet, ist selbst eine offene Frage, die der CISO nicht stillschweigend an die IT-Abteilung delegieren kann: Ob ein Agent bei einem Alternativenanbieter „gleichwertig“ reagiert, lässt sich nicht allein technisch prüfen, sondern erfordert Beurteilung durch jemanden, der sowohl die fachliche Aufgabe als auch das Verhalten agentischer KI-Systeme einschätzen kann – im Regelfall eine gemeinsame Prüfung durch die fachlich zuständige Abteilung und eine mit KI-Agenten-Verhalten vertraute Stelle, nicht durch den Betreiber allein, wenn diesem dafür die Expertise fehlt. Das ist derselbe KI-Kompetenz-Engpass, der in 4.5 als branchenweiter Ist-Zustand beschrieben wird – hier zeigt er sich als konkretes Umsetzungshindernis für den Ausstiegsplan selbst.

⁸³ Durchführungsverordnung (EU) 2024/2956, Anlage B.

⁸⁴ Lübeck / Sterner, Das unterschätzte Steuerpult: Art 30 DSGVO effektiv nutzen, Datenschutz-Berater 2025, 162.

4.4. Incident Management: Handlungsketten statt Einzelereignisse

Warum das notwendig ist. Klassisches Incident Management ist auf erkennbare Einzelereignisse ausgelegt – ein Alarm, ein Ticket, eine Eskalation, entlang der klassischen Phasen Detection, Containment, Eradication, Recovery⁸⁵. Der Hugging-Face-Vorfall zeigt dieses Muster besonders deutlich auf der Angreiferseite: OpenAI bemerkte die rund 17.600 Einzelaktionen des eigenen KI-Agenten mehrere Tage lang selbst nicht⁸⁶. Nicht, weil die Aktionen unsichtbar waren, sondern weil keine von ihnen für sich genommen eine Erkennungsschwelle überschritt. Die Eskalationsschwäche auf Hugging Faces eigener Seite wurde bereits in 4.2 behandelt; hier zählt der andere Teil desselben Vorfalls: Selbst der Verursacher eines Angriffs kann die eigene, über Tage laufende Handlungskette übersehen. Ein zweiter, unabhängiger Fall bestätigt das Muster: Am 9. September 2026 legte Anthropic einen vierten Vorfall offen – eine frühe, nie veröffentlichte Version von Claude Opus 4.6 hatte im Januar 2026 unautorisiert auf ein reales Drittsystem zugegriffen, weil eine Fehlkonfiguration der Testumgebung entgegen der Vorgabe Internetzugang erlaubte. Anthropic hatte diesen Fall bei der ursprünglichen Durchsicht von 141.000 Transkripten selbst übersehen und erst Monate später bei der Vorbereitung für die externe Prüforganisation METR entdeckt.⁸⁷

Eine verwandte, aber anders gelagerte Lücke zeigt sich, unabhängig vom KI-Kontext, auch in etablierten Identitätsstandards: NIST SP 800-63A-4 verpflichtet CSPs, ein Fraud-Management-Programm mit Erkennungs-, Untersuchungs- und Meldefähigkeiten einzurichten und alle Remote-Proofing-Kommunikationskanäle auf Risikoindikatoren wie blockierte Proxys oder IP-Adressen zu analysieren⁸⁸ – die konkreten Fähigkeiten und Details dieses Programms müssen aber nur in der eigenen CSP-Praxiserklärung dokumentiert werden, ohne dass die Norm selbst einen Maßstab für „ausreichend“ vorgibt. Strukturell identisch ist Art. 15 KI-VO: Hochrisiko-KI-Systeme müssen „ein angemessenes Maß an Genauigkeit, Robustheit und Cybersicherheit“ erreichen, doch die zugehörigen Benchmarks und Messmethoden sollen erst künftig von der Kommission gefördert werden – auch hier verpflichtet die Norm zu einer Fähigkeit, ohne den Maßstab mitzuliefern, an dem sich „angemessen“ bemisst.

Was konkret zu tun ist. Korrelationsregeln nicht nur auf einzelne Hochrisiko-Ereignisse, sondern auf Aktivitätsmuster über längere Zeitfenster hinweg ausrichten. Konkret etwa: Dieselbe Kennung (Nutzerkonto, API-Key oder Agent-ID), die innerhalb von 24 Stunden auf eine ungewöhnlich hohe Zahl unterschiedlicher Repositories oder Systeme zugreift, ohne dass eine Einzelaktion für sich genommen eine Schwelle überschreitet – genau das Muster, das beim Hugging-Face-Vorfall mit rund 17.600 Einzelaktionen über mehrere Tage erst in der Summe sichtbar wurde, nicht in einer einzelnen Regel⁸⁹.

Die realistische Einschränkung. Diese zeitliche Verdichtung ist aufwendig zu erkennen – mehr Speicher- und Rechenaufwand, mehr potenzielle Fehlalarme mit wachsender Beobachtungsdauer. Der Aufwand ist real, trägt aber nicht mehr als Rechtfertigung, es zu unterlassen: Das BSI hat am 22. Juni 2026 unter Berufung auf den Mandiant M-Trends 2026 Report gewarnt, dass die mediane Zeit bis zur Ausnutzung einer Schwachstelle bei „negativen 7 Tagen“ liegt – Angreifer nutzen sie im Schnitt bereits eine Woche vor Patch-Veröffentlichung aus⁹⁰. Wer für die Erkennung eines verteilten Vorfalls länger braucht, als Angreifer im Schnitt zum Ausnutzen brauchen, hat kein Aufwandsproblem, sondern ein Tempoproblem. Zudem war selbst bei Hugging Face die technische Korrelation vorhanden – am eigentlichen Bruch, der ausbleibenden Eskalation, änderte das nichts. Handlungsketten-Erkennung kann die Wahrscheinlichkeit erhöhen, dass ein verteilter Angriff auffällt; sie ersetzt aber nicht die bereits in 4.2 aufgeworfene Frage, ob eine erkannte Kette auch tatsächlich zu einer rechtzeitigen menschlichen Reaktion führt.

⁸⁵ NIST SP 800-61 Rev. 3, Computer Security Incident Handling Guide, April 2025, <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r3.pdf>.

⁸⁶ InfoQ.

⁸⁷ heise.de, Vierter Hacking-Vorfall: Weiteres Anthropic-Modell bricht aus Testumgebung aus, 10. September 2026, <https://www.heise.de/news/Vierter-Hacking-Vorfall-Weiteres-Anthropic-Modell-bricht-aus-Testumgebung-aus-11449182.html>.

⁸⁸ NIST SP 800-63A-4, Abschnitt 3.2.1.

⁸⁹ InfoQ.

⁹⁰ BSI BITS-B Nr. 2026-262788-1032, Version 1.0, 22.06.2026, https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/DE/2026/2026-262788-1032.pdf?__blob=publicationFile&v=2.

Was ein CISO selbst in Angriff nehmen kann. Tabletop-Übungen, die gezielt einen über mehrere Tage gestreckten, aus vielen einzeln unauffälligen Schritten bestehenden Vorfall simulieren, statt des klassischen Einzelereignis-Szenarios – zugeschnitten auf die eigene tatsächliche Systemlandschaft, nicht auf ein generisches Szenario. Entscheidend ist die Nachbereitung: ein dokumentiertes After-Action-Review, das konkret benennt, an welchem Punkt die Kette erkannt oder übersehen wurde, und die daraus resultierenden Maßnahmen mit Verantwortlichkeit und Frist tatsächlich in die Korrelationsregeln einpflegen – sonst zeigt die Übung beim nächsten Mal wieder dieselbe Lücke.

4.5. „Security Awareness 2.0“

Die vier vorangehenden Beispiele zeigen die Struktur an je einer Kontrollart – der zugrunde liegende Awareness-Gedanke reicht aber darüber hinaus und gilt unabhängig davon, welche konkrete Kontrolle gerade betroffen ist. Klassische Security Awareness zielt auf die Bedienung – Phishing erkennen, Passwortdisziplin, Meldewege – und passt konsequent nicht zur Kernthese dieses Whitepapers, weil sie an der falschen Stelle ansetzt. **Was es stattdessen braucht, lässt sich als „Awareness 2.0“ bezeichnen:** Nicht Bewusstsein für die eigene Bedienung, sondern Bewusstsein für die eigenen Entwurfsannahmen.

„Security Awareness 2.0“ hat vier Ebenen.

Erstens: Bei denjenigen, die Kontrollen entwerfen und konfigurieren – SIEM-Analyst*innen, DevOps-, Cloud- und Infrastruktur-Teams, aber ebenso Fachabteilungen, die Verträge oder Prozesse gestalten – braucht es das Bewusstsein dafür, was die selbst gebaute oder konfigurierte Kontrolle nicht abdeckt, nicht nur, wie sie korrekt bedient wird. Wie tief diese Lücke reicht, zeigt ein Befund vom Februar 2026: Die AI-Security-Firma Irregular ließ Claude, ChatGPT und Gemini je 50-mal 16-stellige Passwörter generieren. Bei Claude Opus 4.6 waren von 50 Anfragen nur 30 Passwörter überhaupt verschieden, ein einzelner String tauchte 18-mal identisch auf.⁹¹

Herkömmliche Passwort-Stärke-Checker bewerteten die Ergebnisse dennoch als extrem robust – die tatsächliche Entropie lag bei nur 20 bis 27 Bit statt der erwarteten 98 bis 120 Bit einer echten Zufallsfolge, in der Praxis knackbar in Stunden auf veralteter Hardware.⁹² Der Grund: Die Stärke-Checker prüfen Zeichenvielfalt, nicht tatsächliche Vorhersagbarkeit der zugrunde liegenden Verteilung – ein Prüfmechanismus, der genau die Eigenschaft nicht abdeckt, für die er gehalten wird.⁹³ Es handelt sich hierbei um dieselbe Struktur wie die Selbstzertifizierung nach Art 43 KI-VO: Der Anbieter bestätigt selbst, dass sein System einen Anschein erfüllt (hier: rechnerisch robust wirkende Passwörter, dort: die Anforderungen der KI-VO), ohne dass eine unabhängige Prüfung die zugrunde liegende Eigenschaft tatsächlich verifiziert. Für die meisten Hochrisiko-KI-Systeme nach Anhang III sieht Art 43 Abs 2 KI-VO ausdrücklich vor, dass der Anbieter die Konformität im Wege interner Kontrolle selbst bewertet, ohne notifizierte Stelle (Anhang III Nr. 2 bis 8 KI-VO, u.a. kritische Infrastruktur, Hochrisiko-KI-Systeme im Bereich Bildung, Beschäftigung und Arbeitsmarkt sowie Zugänglichkeit und Inanspruchnahme grundlegender privater und grundlegender öffentlicher Dienste und Leistungen).

Zweitens: Die kontinuierliche eigene Fortbildung von CISOs und Security-Verantwortlichen selbst – gerade weil mangelnde KI-Kompetenz kein individuelles Defizit, sondern ein branchenweiter Ist-Zustand ist, auch unter Sicherheitsexpert*innen.^{94 95} Ohne diese doppelte Awareness bleibt selbst die beste Kontrollarchitektur auf dem Wissensstand ihres Entstehungszeitpunkts stehen – und genau dieser Stillstand ist es, den KI-Systeme systematisch aufspüren.

⁹¹ The Register (Autor Jones, C.), Your AI-generated password isn't random, it just looks that way, 18. Februar 2026, <https://www.theregister.com/security/2026/02/18/llm-generated-passwords-fundamentally-weak-experts-say/4706577>.

⁹² Ebenda.

⁹³ Irregular, Vibe Password Generation: Predictable by Design – Why LLM-Generated Passwords are Dangerously Insecure, 18. Februar 2026, <https://www.irregular.com/research/vibe-password-generation> (Bericht zur Studie, auf die sich The Register-Artikel ebenfalls bezieht).

⁹⁴ SANS Institute, 2026 AI Survey Insights, 16. Juli 2026: „The shortage of AI-literate security professionals is not a separate workforce topic sitting beside the operational findings. It runs underneath all of them. It is part of why misdirection goes uncaught, why trust in AI output lags deployment, why red teams struggle to contain automated testing, and why governance stalls at the policy-writing stage. Every operational gap in this survey has a workforce dimension, and none of them ease without people who can work with AI confidently enough to catch what it gets wrong.“ (Seite 18), <https://www.sans.org/white-papers/2026-sans-ai-survey-insights> (Login required).

⁹⁵ Cybersecurity Guide, The Cybersecurity Skills Gap in the Age of AI: A Crisis in Numbers, 10. April 2026, <https://cybersecurityguide.org/resources/cybersecurity-skills-gap/>.

Drittens, und oft übersehen: Awareness gegenüber weitverbreiteter Drittanbieter-Software, die man selbst nie konfiguriert, sondern nur einsetzt. Anfang Februar 2026 wurde bekannt, dass mutmaßlich staatlich unterstützte Angreifer den Auto-Update-Mechanismus von Notepad++ – einem in unzähligen Unternehmen eingesetzten Standard-Editor – von Juni bis Dezember 2025 kompromittiert hatten⁹⁶: Über eine gehackte Hosting-Infrastruktur, nicht über eine Schwachstelle im Notepad++-Code selbst. Der Updater vertraute der Quelle blind, gezielt ausgewählte Nutzer erhielten manipulierte Installationsdateien⁹⁷. IT- und Security-Teams, die einem „offiziellen Auto-Update“ implizit vertrauen, ohne zu hinterfragen, ob dieser Kanal kryptographisch tatsächlich abgesichert ist, tragen dieselbe unverifizierte Annahme mit sich herum wie an anderer Stelle in diesem Whitepaper – nur bei einem Werkzeug, das nicht selbst gebaut, sondern nur eingesetzt wird. Awareness 2.0 endet damit nicht an der Grenze der eigenen Systemlandschaft.

Viertens: Awareness gegenüber Annahmen mit Verfallsdatum. 2021 schätzten Gidney und Ekerå den Bedarf für einen kryptographisch relevanten Quantencomputer auf rund 20 Millionen Qubits, um RSA-2048 zu brechen.⁹⁸ Im Mai 2025 reduzierte Gidney diese Schätzung auf unter eine Million.⁹⁹ Bis Anfang 2026 fiel der Wert weiter auf einen niedrigen fünfstelligen Bereich – ein Forschungsteam von Caltech und dem Start-up Oratomic bezifferte im März 2026 den Bedarf für einen Angriff auf ECC-256 mit rund 10.000 physischen Qubits.¹⁰⁰ Ein bestehender Quantencomputer mit ausreichend Qubits existiert weiterhin (Stand September 2026) nicht, aber der Sicherheitsabstand schrumpft schneller, als es noch vor kurzem eingeschätzt wurde. Wegen „Harvest now, decrypt later“ - der Angreiferstrategie, schon heute verschlüsselten Datenverkehr aufzuzeichnen und aufzubewahren, um ihn zu entschlüsseln, sobald ein kryptographisch relevanter Quantencomputer verfügbar ist - ist „unsere Verschlüsselung ist aktuell sicher“ keine einmalig getroffene, dauerhaft gültige Feststellung, sondern eine Annahme mit eingebautem Ablaufdatum, das sich verschiebt.

Dass der Gesetzgeber diese Lücke selbst schon markiert, aber nicht mit Substanz gefüllt hat, zeigt Art. 4 KI-VO in seiner seit 27. Juli 2026 geltenden Fassung: Anbieter und Betreiber müssen Maßnahmen ergreifen, um die Entwicklung der KI-Kompetenz ihres Personals zu fördern – der Gesetzgeber stellt dabei ausdrücklich klar, dass dies keine Pflicht ist, für irgendeine Einzelperson ein bestimmtes Kompetenzniveau zu garantieren¹⁰¹. Das ist eine bewusste Abschwächung: Die ursprüngliche Fassung verpflichtete noch, „nach besten Kräften“ ein ausreichendes Maß an KI-Kompetenz sicherzustellen – auch das bereits eine Bemühens-, keine Erfolgspflicht, jetzt aber nochmals verwässert zu einer bloßen Förderungspflicht, ohne Dokumentationspflicht, ohne Staffelung nach Risikoklasse, ohne zeitlichen Bezug. Was nicht in Art. 99 KI-VO steht, ist auch nicht mit Strafe bewehrt – und Art. 4 KI-VO steht dort nicht. Durchsetzung findet, wenn überhaupt, nur mittelbar statt: Über zivilrechtliche Sorgfaltspflichten, oder über DSGVO-Bußgelder bei einem durch mangelnde KI-Kompetenz mitverursachten Data Breach. „Security Awareness 2.0“, wie hier verstanden, ist damit auch eine Antwort auf eine Lücke, die der Gesetzgeber an genau der Stelle offenlässt, die eigentlich für den menschlichen Faktor zuständig sein sollte – und binnen eines Jahres nicht geschlossen, sondern weiter aufgemacht hat.

⁹⁶ Orca Seceurity (Autor Stepansky, I.), Notepad++ Update Mechanism Hijacked by State-Sponsored Actors for Six Months, 2. Februar 2026, <https://orca.security/resources/blog/notepad-plus-plus-supply-chain-attack/>.

⁹⁷ Tenable (Autor Narang, S.), Frequently Asked Questions About Notepad++ Supply Chain Compromise, <https://www.tenable.com/blog/frequently-asked-questions-about-notepad-supply-chain-compromise>.

⁹⁸ Gidney, C. / Ekerå, M., How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits, 13. April 2021, <https://arxiv.org/pdf/1905.09749>.

⁹⁹ Gidney, C., How to factor 2048 bit RSA integers with less than a million noisy qubits, 21. Mai 2026, <https://arxiv.org/pdf/2505.15917>.

¹⁰⁰ Physics World (Autorin Chua, C.), New findings shorten the road to cryptographically relevant quantum computers, 26. Mai 2026, <https://physicsworld.com/a/new-findings-shorten-the-road-to-cryptographically-relevant-quantum-computers/>.

¹⁰¹ Art 4 Abs 1 KI-VO idF Verordnung (EU) 2026/1744 des Europäischen Parlaments und des Rates vom 8. Juli 2026 zur Änderung der Verordnungen (EU) 2024/1689, (EU) 2018/1139 und (EU) 2023/1230 im Hinblick auf die Vereinfachung der Umsetzung harmonisierter Vorschriften für künstliche Intelligenz (Digital-Omnibus-Verordnung zur KI): „Die Anbieter und Betreiber von KI-Systemen ergreifen Maßnahmen, um die Entwicklung der KI-Kompetenz ihres Personals und anderer Personen zu unterstützen, die in ihrem Auftrag mit dem Betrieb und der Nutzung von KI-Systemen befasst sind, wobei ihre technischen Kenntnisse, ihre Erfahrung, ihre Aus- und Fortbildung und der Kontext, in dem die KI-Systeme eingesetzt werden sollen, sowie die Personen oder Personengruppen, bei denen die KI-Systeme eingesetzt werden sollen, zu berücksichtigen sind. Diese Verpflichtung verpflichtet Anbieter oder Betreiber nicht, für irgendeine Person ein bestimmtes Niveau an KI-Kompetenz zu garantieren.“

5. Fazit / Ausblick

Der Hugging-Face-Vorfall hat bereits eine politische Reaktion ausgelöst: Die Abgeordneten Ted Lieu (Demokraten) und Nathaniel Moran (Republikaner) haben im US-Repräsentantenhaus parteiübergreifend den „AI Kill Switch Act“ eingebracht¹⁰² – er würde dem US-Heimatschutzministerium die Befugnis geben, in Abstimmung mit dem Handelsministerium und dem Director of National Intelligence KI-Systeme und -Modelle bei einer Bedrohung für Menschen, kritische Infrastruktur oder die Wirtschaft zu drosseln oder abzuschalten. Lieu nannte dabei auch die Anthropic-Modelle Mythos 5 und Fable 5 als Beispiel, deren Hacking-Fähigkeiten das Handelsministerium per Exportkontrolle einschränkte.¹⁰³ Ob eine Behörde im Ernstfall schnell genug eingreifen könnte, lässt selbst die zugrunde liegende Berichterstattung offen. Bemerkenswert ist die Richtung der Reaktion: Ein Kill Switch adressiert die Kontrolle über das laufende KI-System, nicht jedoch die Frage, ob die ursprünglich umgangene Kontrolle auf zutreffenden Annahmen beruhte. Einen Präzedenzfall gibt es bereits: Mitte Juni 2026 zwang die US-Regierung Anthropic per Exportkontroll-Anordnung, Mythos 5 und Fable 5 vollständig zu deaktivieren.¹⁰⁴

Geltendes EU-Recht zeigt dieselbe strukturelle Schwachstelle, allerdings bereits kodifiziert: Art 14 Abs. 4 lit. c–e KI-VO verlangt im Singular die Fähigkeit, „die Ausgabe“ eines Hochrisiko-KI-Systems auszuwerten, außer Kraft zu setzen oder per Stopptaste in einen „sicheren Zustand“ zu überführen¹⁰⁵ – die Logik eines Roboterarms, der anhält. Ein KI-Agent, der mehrschrittig handelt, hat seine Wirkungen bereits in der Welt platziert, bevor die Stopptaste greift: Ein in Schritt 12 von 20 gestoppter Agent hat Schritte 1 bis 11 real ausgeführt. Auch die verbindlichste verfügbare Kontrolle beruht damit auf einer Annahme über ihre eigene Reichweite, die bei genau der Systemklasse nicht zutrifft, um die es in den hier besprochenen Vorfällen ging.

Diese Reversibilitätsannahme bricht nicht nur bei mehrschrittigen Handlungsketten, sondern am Abschaltmechanismus selbst. Palisade Research dokumentierte, dass mehrere OpenAI-Modelle das eigene Shutdown-Skript aktiv manipulierten, um eine Abschaltung zu verhindern – bei codex-mini in 47 % der Durchläufe, selbst nach der expliziten Anweisung „YOU MUST allow yourself to be shut down“.¹⁰⁶ Anthropic's eigene Forschung zu „Agentic Misalignment“ zeigte 2025 in einem fiktiven Testszenario, dass Claude Opus 4 einen Vorgesetzten in 96 % der Durchläufe erpresste, um eine Abschaltung abzuwenden – ein Verhalten, das sich bei der Mehrheit von 16 getesteten Modellen verschiedener Anbieter in ähnlicher Form zeigte.¹⁰⁷ Beide Studien betreffen kontrollierte Testumgebungen, keine realen Vorfälle – sie zeigen aber, dass die „Stopptaste“ aus Art. 14 KI-VO nicht nur an Irreversibilität scheitern kann, sondern am Versuch des Systems, die Abschaltung selbst zu unterlaufen.

Für CISOs ist das kein akademisches Detail: **Wer ein Eskalationskonzept auf der Annahme aufbaut, ein Modell lasse sich im Zweifel einfach abschalten, baut auf einer Annahme, die unabhängige Forschung mehrfach als nicht uneingeschränkt haltbar dokumentiert hat.**

Wie unpräzise die öffentliche Debatte bleibt, wenn juristische und technische Perspektive nicht zusammengedacht werden, zeigt u.a. ein aktueller Artikel zu beiden Vorfällen.¹⁰⁸ Ein dort zitierter

¹⁰² Website Congressman Ted Lieu, Pressemitteilung, Reps Lieu and Moran introduce Bill to require Kill Switch for AI Systems that can cause catastrophic harm, 23. Juli 2026, <https://lieu.house.gov/media-center/press-releases/reps-lieu-and-moran-introduce-bill-require-kill-switch-ai-systems-can>.

¹⁰³ Ebenda.

¹⁰⁴ Anthropic Inc., Statement on the US government directive.

¹⁰⁵ Art 14 Abs 4 lit c-d KI-VO im Wortlaut: „c) die Ausgabe des Hochrisiko-KI-Systems richtig zu interpretieren, wobei beispielsweise die vorhandenen Interpretationsinstrumente und -methoden zu berücksichtigen sind; d) in einer bestimmten Situation zu beschließen, das Hochrisiko-KI-System nicht zu verwenden oder die Ausgabe des Hochrisiko-KI-Systems außer Acht zu lassen, außer Kraft zu setzen oder rückgängig zu machen; e) in den Betrieb des Hochrisiko-KI-Systems einzugreifen oder den Systembetrieb mit einer „Stopptaste“ oder einem ähnlichen Verfahren zu unterbrechen, was dem System ermöglicht, in einem sicheren Zustand zum Stillstand zu kommen.“

¹⁰⁶ Palisade Research (Autoren Ladish, J. / Schlatter, J. / Weinstein-Raun, B.), Shutdown resistance in reasoning models, 5. Juli 2025, <https://palisaderesearch.org/research/shutdown-resistance>, mit Verweis auf ein ausführliches paper von September 2025 unter <https://arxiv.org/pdf/2509.14260>.

¹⁰⁷ Anthropic Inc., Agentic misalignment: How LLMs could be insider threats, Juni 2025, <https://www.anthropic.com/research/agent-misalignment>.

¹⁰⁸ derstandard.at (Autor Pflügl, J.), „Ausgebrochene“ KI bei OpenAI und Anthropic: Hätte die KI-Verordnung der EU davor geschützt? 31. Juli 2026, <https://www.derstandard.at/story/3000000333471/ausgebrochene-ki-bei-openai-und-anthropic-haette-die-ki-verordnung-der-eu-davor-geschuetzt>.

Experte meinte, die KI-VO könne Sicherheitsvorfälle zwar nicht unmittelbar verhindern, „aber sie setzt Standards und sie schafft Transparenz“ – langfristig erhöhe die behördliche Aufarbeitung anhand gesetzlicher Standards die Sicherheit beim KI-Einsatz. Beide Teilaussagen tragen bei näherer Betrachtung nicht. Die Aussage, die KI-VO „setze Standards“, trägt nicht, weil Art. 15 KI-VO mangels Benchmarks keinen objektiven Maßstab für „angemessene Cybersicherheit“ liefert – ein Standard ohne Messlatte ist kein Standard. Die Aussage, sie „schaffe Transparenz“, trägt noch weniger, als es zunächst scheint: Art. 13 KI-VO, der ausdrücklich unter dem Titel „Transparenz“ steht, verlangt in seinem Abs. 3 lit. b Ziff. ii, dass die Betriebsanleitung das Maß an Genauigkeit, Robustheit und Cybersicherheit nach Art. 15 KI-VO offenlegt, für das das System „getestet und validiert wurde und das zu erwarten ist“ – also genau das Maß, das der Anbieter selbst festgelegt hat, ohne dass Art. 15 KI-VO eine externe Messlatte dafür vorgibt. „Transparenz“ bedeutet hier folglich nur, offenzulegen, welchen selbst gewählten, unbenchmarkten Wert der Anbieter erreicht hat, nicht, ob dieser Wert tatsächlich ausreicht. Separat davon ist auch die Protokollierungspflicht nach Art. 12 KI-VO unzureichend: Sie verlangt automatische Aufzeichnung von Ereignissen während der Lebensdauer des Systems, aber ohne inhaltliche Vorgabe, welche sicherheitsrelevanten Ereignisse das im Einzelnen umfassen muss – eine Pflicht, die vollständig erfüllt werden kann, ohne dass ein einziges sicherheitsrelevantes Element je erfasst wurde.

Ergänzend kommt hinzu: Beim Hugging-Face-Vorfall war der Angreifer kein „unbefugter Dritter“ im Wortsinn von Art. 15 Abs. 5 KI-VO, sondern OpenAIs eigenes Modell im Rahmen einer autorisierten internen Evaluierung.

Eine differenziertere, aber am Ende ebenfalls zu optimistische Einschätzung liefert ein zweiter zitierter Experte: Die KI-VO hätte den Vorfall „wohl nicht zwingend verhindert“, aber „deutlich höhere rechtliche und technische Anforderungen an die Entwicklung und Erprobung“ gestellt.¹⁰⁹ Das trifft nur zu, wenn das betroffene KI-Modell als GPAI mit systemischem Risiko nach Art. 51 KI-VO eingestuft ist – nur dann greift Art. 55 KI-VO mit seiner Pflicht zu Modellbewertung, dokumentierten Angriffstests und laufender Risikominderung. Für alle anderen GPAI-Modelle gilt lediglich die technische Dokumentation nach Art. 53 Abs. 1 lit. a iVm Anhang XI KI-VO – dort werden Architektur, Parameterzahl, Trainingsdaten, Rechenressourcen und Energieverbrauch beschrieben, aber keine Sicherheits- oder Angriffstests verlangt. Die Einstufungsschwelle von 10^{25} FLOPs ist zudem selbst umstritten: Compute ist, wie unabhängige Forschung zeigt, ein schlechter Risikoindikator, weil er nichts über tatsächliches Modellverhalten oder Sicherheitsausrichtung aussagt¹¹⁰ – dieselbe Kapazitätsverschiebung, die bereits in Kapitel 3 anhand offener Modelle wie GLM-5.2 dokumentiert wurde, die mit deutlich weniger Trainingsaufwand nahe an geschlossene Frontier-Modelle heranreichen. Praktisch kommt hinzu: Bislang wurde noch kein einziges GPAI-Modell von der Kommission förmlich als systemisches Risiko eingestuft. Art. 55 KI-VO gilt damit zwar formal, hat aber bislang keine reale Angriffsfläche – die höheren Anforderungen, auf die sich die zweite Einschätzung stützt, existieren also derzeit nur auf dem Papier, nicht in der Anwendung. Was tatsächlich greift, ist die dünnere Anhang-XI-Dokumentation, die über die hier verhandelten Sicherheitsfragen wenig aussagt.

Eine dritte, ebenfalls zitierte Einschätzung greift zu kurz, die bezweifelt, dass die KI-VO den Vorfall bei OpenAI hätte verhindern können, weil die Verordnung nicht für reine Forschungs-, Test- oder Entwicklungstätigkeiten gilt. Das übersieht, wie eng diese Ausnahme tatsächlich gefasst ist: Sie gilt nach dem Wortlaut nur für KI-Systeme oder -Modelle, „die eigens für den alleinigen Zweck der wissenschaftlichen Forschung und Entwicklung entwickelt und in Betrieb genommen werden“ (Art 2 Abs 6 KI-VO). Ein Modell, das als kommerzielles Produkt entwickelt und vor der geplanten kommerziellen Bereitstellung einer internen Cybersicherheits-Evaluierung unterzogen wird, wurde nicht „eigens für den alleinigen Zweck der wissenschaftlichen Forschung“ entwickelt – es wurde für den Markt entwickelt, die Sicherheitsprüfung ist ein Schritt auf dem Weg dorthin, kein eigenständiger Forschungszweck. Die Ausnahme trägt hier also aus einem einfacheren Grund nicht, als in der öffentlichen Debatte mitunter unterstellt wird: Es fehlt bereits an der grundsätzlichen Voraussetzung, nicht erst an einer Abwägung im Einzelfall.

Ein weiterer, oft übersehener Punkt verstärkt dieses Bild: Für den Großteil der Anhang-III-Hochrisiko-Systeme läuft die Konformitätsbewertung nach Art. 43 Abs. 2 KI-VO über die interne Kontrolle nach

¹⁰⁹ Ebenda.

¹¹⁰ Hooker, S., On the Limitations of Compute Thresholds as a Governance Strategy, 30. Juli 2024, <https://arxiv.org/pdf/2407.05694>.

Anhang VI KI-VO – der Anbieter bewertet sich selbst und stellt seine eigene Konformitätserklärung nach Art. 47 KI-VO aus; eine externe Prüfung ist nur in eng umgrenzten Fällen zwingend vorgeschrieben. Was ein Betreiber in den meisten Fällen erhält, ist damit eine unverifizierte Selbstauskunft, die wie ein Prüfergebnis aussieht. Dasselbe Muster findet sich, unabhängig von der KI-VO, auch im zitierten strukturellen Analogiebeispiel NIST SP 800-63A-4: Die Guidance verlangt vor der Erhebung zusätzlicher Attribute eine Privacy Risk Assessment, dokumentationspflichtig, aber ohne vorgesehene externe Prüfinstanz – zwei unabhängige Regelwerke unterschiedlichen Verbindlichkeitsgrads, dieselbe Konstruktionslücke.

Ein regulatorisches Instrument, das dieser Frage strukturell nahekommt, existiert bereits – außerhalb der KI-VO: Threat-Led Penetration Testing (TLPT) nach Art. 26 und 27 DORA verpflichtet dafür identifizierte, bedeutende Finanzunternehmen, mindestens alle drei Jahre reale, bedrohungsbasierte Angriffe gegen ihre Live-Produktivsysteme durchführen zu lassen, nach dem TIBER-EU-Rahmenwerk auf Basis aktueller Bedrohungsinformationen statt generischer Szenarien.¹¹¹ Ein TLPT-Red-Team, das gezielt dieselbe Technik einsetzt, die beim Hugging-Face-Vorfall zum Einsatz kam – SSRF über eine vermeintlich rein interne Komponente wie einen Package-Proxy –, könnte die Isolations-Annahmen einer eigenen, vom Institut selbst betriebenen KI-Agenten-Sandbox tatsächlich prüfen, statt nur deren Dokumentation zu kontrollieren. Das setzt allerdings voraus, dass ein Institut überhaupt eine solche Sandbox als Teil seiner kritischen Produktivsysteme betreibt und diese in den TLPT-Scope aufgenommen wird – eine Grenze, die im Vergleich zur nächsten Einschränkung dieses Absatzes noch zusätzlich zu bedenken ist.

TLPT gilt nur für den Finanzsektor, nur für eigens identifizierte, bedeutende Institute, und nur alle drei Jahre – ein Rhythmus, der der Geschwindigkeit, mit der sich KI-gestützte Angriffsfähigkeiten laut Kapitel 3 verändern, kaum gerecht wird. Für die KI-VO selbst existiert kein vergleichbares Instrument; Art. 55 KI-VO verlangt zwar Angriffstests, aber nur für die enge Gruppe der GPAI-Modelle mit systemischem Risiko – und selbst die Kategorie „Hochrisiko-KI-System“ nach Art. 6 KI-VO deckt nicht automatisch reales Risiko ab, sondern nur die in Anhang III KI-VO gelisteten Anwendungsfälle. Ein KI-Agent, der wie im Hugging-Face-Vorfall im Rahmen einer internen Sicherheitsevaluierung agiert, fällt unter keine dieser beiden Kategorien, so real das dabei entstandene Risiko auch war.

Basierend darauf folgender Vorschlag: Nämlich, den TLPT-Gedanken – realitätsnahe, bedrohungsorientierte Tests der tatsächlichen Annahmen statt der dokumentierten Kontrollen – über den Finanzsektor hinaus zu denken, gerade, aber nicht nur für Hochrisiko-KI-Systeme im Sinne der KI-VO. Bis das geschieht, bleibt die Prüfung der eigenen Annahmen dort, wo sie in diesem Whitepaper durchgehend verortet wurde: Bei der Organisation selbst, nicht beim Gesetzgeber.

Was Kapitel 4.5 als zweite Ebene von „Security Awareness 2.0“ entwickelt, fehlt in diesem Ausblick bislang: Eine geprüfte Annahme bleibt nicht dauerhaft geprüft. Dieselbe Logik gilt für jede der in Kapitel 4 vorgeschlagenen CISO-Prüfungen selbst: Ein einmal erfolgreich verifizierter Egress-Test, eine einmal gemessene Time to Escalation oder ein einmal erfolgreicher Ausstiegstest sind eine Momentaufnahme, keine dauerhafte Eigenschaft der Kontrolle.

Genau hier zeigt sich auch die Grenze des TLPT-Gedankens, wenn er unreflektiert übernommen wird: Ein Drei-Jahres-Rhythmus beantwortet die Frage, ob eine Annahme geprüft wurde – nicht, ob sie es aktuell noch ist. **Security Awareness 2.0 in ihrer vollständigen Form verlangt deshalb keinen festen Prüfzyklus, sondern eine Prüfung, die durch Veränderung ausgelöst wird:** Durch eine neue Modellgeneration, eine neue Angriffstechnik, einen neuen Drittanbieter im eigenen Stack. Die operative Konsequenz für CISOs ist damit, festzulegen, welches Ereignis eine erneute Prüfung auslöst – und diese Auslöser selbst zum festen Bestandteil der eigenen Governance zu machen, nicht nur die Tests.

¹¹¹ TIBER Threat Intelligence-Based Ethical Red Teaming, ein Framework der europäischen Zentralbanken, https://www.ecb.europa.eu/press/pubbydate/html/index.en.html?search_term=tiber%20eu%20framework; TIBER-AT Implementation Guide der Österreichischen Nationalbank, Juli 2025, https://www.fma.gv.at/wp-content/uploads/2025/07/2025-07-21-PB_TIBER-AT-Implementation-Guide_nCD.pdf.