The Oversight Illusion

Wie Art. 14 KI-VO eine Sicherheitslücke vorschreibt, die OWASP längst als Angriffsvektor kennt

Bettina Sterner, BA BA
kaffeekipferl.at
Juli 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. Verweise auf Entwürfe, Konsultationsfassungen oder noch nicht endgültig angenommene Leitlinien und Normen sind als solche gekennzeichnet und stehen unter dem Vorbehalt der jeweils aktuellen Fassung.

Deutsch

Der Gesetzgeber schreibt in DSGVO, KI-VO und CRA eine Schwachstelle vor: Menschliche Kontrolle gilt in allen drei Regimen als Schutzmaßnahme — nicht als Ziel. Genau das macht sie angreifbar. Das OWASP Top 10 for Agentic Applications 2026 benennt die gezielte Ausnutzung dieses Vertrauens (Human-Agent Trust Exploitation, ASI09) als eigenständigen, aktiv einsetzbaren Angriffsvektor: Angreifer können Systemausgaben so gestalten, dass sie Vertrauen erzeugen und dadurch inhaltliche Prüfung gezielt umgehen.

Automation Bias ist empirisch robust belegt, und die EuGH-Rechtsprechung zu Art. 22 DSGVO hat das Problem der Scheinkontrolle bereits vor der KI-VO erkannt. Compliance mit diesen Normen erzeugt damit nicht Sicherheit, sondern eine neue, vorhersehbare Angriffsfläche — exakt dort, wo der Rechtsrahmen sie am wenigsten vermutet.

English

The legislator mandates a vulnerability in GDPR, the AI Act, and the CRA: Human control counts as a safeguard in all three regimes — not as a target. That is exactly what makes it attackable. The OWASP Top 10 for Agentic Applications 2026 identifies the deliberate exploitation of this trust (Human-Agent Trust Exploitation, ASI09) as a distinct, actively deployable attack vector: attackers can shape system outputs to build trust and thereby deliberately bypass substantive review.

Automation bias is empirically well documented, and CJEU case law on Article 22 GDPR flagged the problem of sham control before the AI Act existed. Compliance with these norms therefore does not produce safety, but a new, foreseeable attack surface — exactly where the legal framework least expects it.

Auf einen Blick

- **Die These:** Der Gesetzgeber macht menschliche Aufsicht zur Schutzmaßnahme — nicht zum Ziel. Genau das macht sie angreifbar. Was OWASP als aktiv ausnutzbaren Angriffsvektor klassifiziert, stuft der CRA parallel als meldepflichtige Produktschwachstelle ein: Aus einer Compliance-Frage wird ein Sicherheitsvorfall mit eigenem Fristenregime.
- **Die fehlende Infrastruktur:** Eine Aufsichtsperson kann nur erkennen, was von der Norm abweicht, wenn sie weiß, was normal ist. Genau diese Vergleichsgrundlage liefert die KI-VO an keiner Stelle — weder als Baseline, noch als verständliche Anleitung, noch als belastbare Leistungsdaten. Wirksame Aufsicht scheitert damit oft schon an der Werkzeugausstattung, nicht am Willen.
- **Die fehlende Verifikation:** Kompetenz wird vorausgesetzt, nie geprüft. Wirksamkeit wird behauptet, nie nachgewiesen. Der Zeitpunkt, an dem ein Eingriff noch etwas bewirkt, ist nirgends definiert. Drei Fiktionen, die zusammen die Kontrollinstanz selbst zum schwächsten Glied machen.
- **Der ignorierte Vorläufer:** Der EuGH hat formale menschliche Beteiligung schon vor der KI-VO als Scheinkontrolle zurückgewiesen — und die KI-VO nimmt darauf an keiner Stelle Bezug, nicht einmal dort, wo sie es am direktesten könnte.
- **Das Eingeständnis im Normtext:** Der Gesetzgeber sagt es selbst, ohne es zu merken — menschliche Aufsicht soll Risiken auffangen, die trotz voller Compliance fortbestehen. Nur: Ob dieses Auffangnetz trägt, wird nirgends verlangt nachzuweisen.

1. Eine Beobachtung, die zum Weiterdenken anregt: „Human controls the loop“

Im Juni 2026 hielt David Rosenthal, Partner bei VISCHER Rechtsanwälte, in Zürich einen Vortrag zu Agentic AI, dessen Kernthese er auch auf LinkedIn online veröffentlichte: „**Human in the Loop**" greift zu kurz. Entscheidend ist, dass der Mensch den Ablauf tatsächlich kontrolliert — „**Human controls the Loop**". Wird diese Kontrolle konsequent umgesetzt, können autonome Agenten in bestimmten Bereichen sachgerecht eingesetzt werden. Wer sich dagegen auf nachträgliche Prüfung von KI-Ergebnissen verlässt, erzeugt ein falsches Sicherheitsgefühl — Rosenthal nennt das Kind beim Namen: **Automation Bias.**

Der Vortrag stand neben Beiträgen zu Sicherheitsfragen und zur betriebswirtschaftlichen Seite von Agentic AI, vor einem Fachpublikum aus Datenschutz- und Compliance-Expertinnen und -Experten. Die Autorin kennt den Vortrag nicht aus eigener Teilnahme, sondern über Rosenthals LinkedIn-Beitrag — genau diese Beobachtung hat sie aufgrund ihrer eigenen praktischen Perspektive sofort beschäftigt.

Rosenthals Unterscheidung führt zu einer konkreten Frage: Wie hat der Gesetzgeber „wirksame menschliche Aufsicht" tatsächlich ausgestaltet — als echte Kontrollfähigkeit oder nur als deren Behauptung? Genau das entscheidet, ob Compliance mit Art. 14 KI-VO tatsächlich Sicherheit schafft oder nur den Anschein davon. Rosenthal präsentiert „Human controls the Loop" dabei als Ansatz, nicht als bereits erwiesen überlegenes Modell. Dieses Whitepaper arbeitet die Schwächen von „Human in the Loop" heraus und kommt zu dem **Schluss, dass auch „Human controls the Loop" nicht trägt — weil die dafür notwendige Grundlage fehlt.**

2. Wenn Schutz zur Schwachstelle wird

Art. 22 Abs. 1 DSGVO gibt betroffenen Personen das Recht, nicht ausschließlich einer automatisierten Entscheidung unterworfen zu werden, die ihr gegenüber eine rechtliche Wirkung entfaltet oder sie erheblich beeinträchtigt. In den Fällen des Art 22 Abs 2 lit a und c DSGVO (erforderlich für die Erfüllung oder den Abschluss eines Vertrages zwischen betroffener Person und Verantwortlichem bzw. ausdrückliche Einwilligung der betroffenen Person) verlangt Art 22 Abs 3 DSGVO das Recht auf Erwirkung des Eingreifens einer Person auf Seiten des Verantwortlichen. Art. 14 Abs 1 KI-VO wiederum verlangt für Hochrisiko-KI-Systeme iSd Art 6 KI-VO, dass sie so gestaltet sind, dass ein Mensch sie während der Dauer der Verwendung wirksam beaufsichtigen kann.

Beide Normen setzen an unterschiedlichen Punkten an — Art. 22 DSGVO nachträglich, auf Antrag der betroffenen Person; Art. 14 KI-VO laufend, während des Betriebs —, verlangen aber beide ausdrücklich *wirksame* menschliche Beteiligung, ohne festzulegen, woran diese Wirksamkeit zu erkennen ist oder wie sie sichergestellt wird. Bei Art. 22 DSGVO musste das erst der EuGH im Nachhinein klären; bei Art. 14 KI-VO fehlt bislang selbst diese nachträgliche Instanz. Der Gerichtshof stellte in SCHUFA (C-634/21) und Dun & Bradstreet (C-203/22) nachträglich klar, dass bloß formale menschliche Beteiligung nicht ausreicht — ein Beleg dafür, dass der Normtext diese Substanzialität selbst nicht erzwingt. Für Art. 14 KI-VO zeigt das die Herleitung in Kapitel 4: Weder liefert Art. 12 KI-VO eine Baseline, gegen die Abweichungen erkennbar wären, noch verlangt Art. 14 Abs. 4 KI-VO eine Verifikation, dass die geforderten Fähigkeiten bei der Aufsichtsperson tatsächlich vorhanden sind.

Automation Bias — die Tendenz, automatisierten Ergebnissen unkritisch zu vertrauen — unterläuft diese Kontrollfähigkeit zusätzlich, unabhängig von der Sorgfalt der einzelnen Person. Skitka, Mosier und Burdick wiesen bereits 1999 nach, dass Testpersonen mit einem zuverlässigen, aber nicht perfekten automatisierten Entscheidungshilfe-System in einer simulierten Überwachungsaufgabe schlechter

abschnitten als Testpersonen ganz ohne Automatisierung[1]. Dass der Effekt mit heutigen KI-Systemen fortbesteht, bestätigt der International AI Safety Report 2026 branchenübergreifend, unter anderem in der medizinischen Diagnostik[2], sowie ein aktuelles systematisches Review von 35 Studien mit knapp 20.000 Teilnehmenden[3]. Ruschemeier und Hondrich kommen zudem zu einem Befund, der die Kernthese dieses Papers stützt: Obwohl menschliche Aufsicht gesetzlich betont wird, adressieren weder DSGVO noch KI-VO Automation Bias in hybriden Mensch-KI-Entscheidungen angemessen[4].

Das OWASP Top 10 for Agentic Applications 2026 geht über die passive Bias-Forschung hinaus. Human-Agent Trust Exploitation (ASI09) beschreibt nicht die Neigung selbst, sondern ihre gezielte Ausnutzung: Ein Angreifer gestaltet Systemausgaben so, dass sie Vertrauen erzeugen und inhaltliche Prüfung gezielt umgehen. OWASP führt das als eigene Kategorie, weil Agentic-AI-Systeme mit Tool-Chains und autonomer Aktionsausführung einen aktiven Angriffsvektor eröffnen — die passive Fehleinschätzung, die klassische Automation-Bias-Forschung untersucht, deckt das nicht ab.

Daraus folgt die Kernthese: DSGVO und KI-VO behandeln menschliche Kontrolle als Schutzmaßnahme, ohne sie gegen genau diese Angriffsform abzusichern. Der CRA verschärft die Konsequenz auf einer dritten Regelungsebene. Fällt ein Hochrisiko-KI-System unter den CRA-Produktbegriff (Art. 3 Nr. 1 CRA), greift überdies die Meldepflicht aus Art. 14 CRA; die zugrunde liegende Beobachtungspflicht für aktiv ausgenutzte Schwachstellen ergibt sich aus Anhang I Teil II Nr. 3 CRA. Human-Agent Trust Exploitation nach ASI09 fällt darunter, sobald sie tatsächlich ausgenutzt wird — Einzelheiten dazu in Kapitel 5. Aus einer Compliance-Frage (Ist Art. 14 KI-VO eingehalten?) wird damit eine produktsicherheitsrechtliche Meldepflicht mit eigenem, deutlich schärferem Fristenregime.

Die folgenden Kapitel leiten diese These im Detail her: Kapitel 3 zeigt an der DSGVO, dass das Muster der Scheinkontrolle nicht neu ist. Kapitel 4 führt sie an der KI-VO selbst durch, Artikel für Artikel. Kapitel 5 zeigt die produktsicherheitsrechtliche Konsequenz über den CRA. Kapitel 6 bestätigt das Ergebnis von der technischen Seite, unabhängig von der juristischen Analyse. Kapitel 7 kehrt zum Ausgangspunkt zurück und zeigt, was die Herleitung für Rosenthals Unterscheidung in der Einleitung dieses Whitepapers bedeutet: Auch „Human controls the loop" trägt nicht, solange die dafür nötige Grundlage fehlt.

3. Der Präzedenzfall - DSGVO

3.1 Formale Beteiligung ist keine Kontrolle

Art. 22 Abs. 1 DSGVO gewährt betroffenen Personen das Recht, keiner ausschließlich automatisierten Entscheidung unterworfen zu werden, die rechtliche Wirkung entfaltet oder sie erheblich beeinträchtigt. Der EuGH legte die Norm in SCHUFA (C-634/21) weit aus: Ein automatisiert erstellter Wahrscheinlichkeitswert fällt bereits dann darunter, wenn ein Dritter — etwa eine Bank — seine Entscheidung „maßgeblich“ darauf stützt. Damit erteilte der Gerichtshof einer verbreiteten Verteidigungsstrategie eine Absage: Ein Prüfschritt beim Dritten, der den Score nur formal zur Kenntnis nimmt, ohne ihn inhaltlich zu hinterfragen oder davon im Einzelfall abzuweichen, reicht nicht, um die automatisierte Verarbeitung aus dem Anwendungsbereich herauszuhalten. Fällt eine Verarbeitung damit unter Art. 22 Abs. 1 DSGVO, stellt sich die Anschlussfrage, was der Verantwortliche der betroffenen Person darüber offenlegen muss.

[1] Skitka, L, Mosier, K; Does automation bias decision-making? November 1999; https://www.researchgate.net/publication/222507469_Does_automation_bias_decision-making.

[2] Herausgeber britische Regierung; Februar 2026; https://arxiv.org/pdf/2602.21012.

[3] Romeo, G., Conti, D.; Exploring automation bias in human-AI collaboration: a review and implications for explainable AI; Juli 2025; https://link.springer.com/article/10.1007/s00146-025-02422-7.

[4] Ruschemeier, H., Hondrich, L.J.; Automation bias in public administration – an interdisciplinary perspective from law and psychology; September 2024; https://www.sciencedirect.com/science/article/pii/S0740624X24000455.

Dun & Bradstreet Austria (C-203/22) präzisiert, worauf sich dieser Auskunftsanspruch nach Art. 15 Abs. 1 lit. h DSGVO konkret erstreckt: Verantwortliche müssen Verfahren und Grundsätze so erläutern, dass die betroffene Person nachvollziehen kann, welche ihrer Daten auf welche Art verwendet wurden — etwa durch die Angabe, welche Abweichung zu einem anderen Ergebnis geführt hätte. Vollständige Offenlegung des Algorithmus verlangt der Gerichtshof nicht, eine Erklärung, die tatsächliche Anfechtung ermöglicht, schon. Geschäftsgeheimnisse rechtfertigen keine pauschale Auskunftsverweigerung: Der EuGH erklärte die entsprechende österreichische Ausnahme (§ 4 Abs. 6 DSG) für unionsrechtswidrig — im Streitfall muss der Verantwortliche die Information der Aufsichtsbehörde oder dem Gericht vorlegen, die dann abwägen.

3.2 Eine Fähigkeit, die der EDSA selbst für hypothetisch hält

Wie wenig das im Regulierungsalltag angekommen ist, zeigt der Europäische Datenschutzausschuss („EDSA“) selbst. Die Guidelines 03/2026 on web scraping in the context of generative AI[5] übertragen den „responsibilities, powers and capabilities“-Test aus GC & Others (C-136/17) auf die unbeabsichtigte Verarbeitung besonderer Datenkategorien beim Web Scraping für KI-Training.

Innerhalb dieses Rahmens verlangt der EDSA unter anderem Widerstandsfähigkeit gegen Privacy-Angriffe nach Stand der Technik— und räumt im nahezu selben Atemzug ein, dass Machine Unlearning, die tatsächliche Fähigkeit, Daten aus einem trainierten Modell zu entfernen, erst „through technological progress" verfügbar werden könnte. Konjunktiv, Zukunftsversprechen, keine bestehende Fähigkeit. Verfügbar bleibt, was auf der Ausgabeebene ansetzt: Output-Filter, Prompt-Restriktionen. Die Modellgewichte selbst bleiben unzugänglich.

Der EDSA fordert damit eine Fähigkeit, deren Existenz er im selben Absatz als hypothetisch einstuft — dasselbe strukturelle Muster wie bei Art. 14 KI-VO, nur eine Regelungsebene früher. Bemerkenswert: Der Test gilt nur für die unbeabsichtigte Verarbeitung, nicht für die generelle Modell-Governance — selbst im engsten, selbst gesetzten Anwendungsfall bleibt die vorausgesetzte Fähigkeit fiktiv.

Was trotz des Titels „in the context of generative AI" fehlt: Jede Verzahnung mit der KI-VO, etwa mit Art. 10 KI-VO (Data Governance) oder Art 11 KI-VO (technische Dokumentation) und Art 12 KI-VO (Aufzeichnungspflichten). Selbst dort, wo der Gesetzgeber integrierte Governance demonstrieren könnte, bleibt der Blick silohaft auf das eigene Regelwerk beschränkt — obwohl die Ursache regelwerksübergreifend liegt.

3.3 Ein Beschwerdemechanismus ohne Maßstab

Art. 27 KI-VO verpflichtet einen begrenzten Kreis von Betreibern — öffentliche Stellen, private Betreiber öffentlicher Dienstleistungen, Betreiber von Kreditwürdigkeits- und Versicherungsrisikosystemen nach Anhang III Nr. 5 lit. b und c KI-VO — zur Grundrechte-Folgenabschätzung. Art. 27 Abs. 1 lit. f KI-VO verlangt dafür die Beschreibung von „Regelungen für die interne Unternehmensführung und Beschwerdemechanismen" — mehr nicht.

Das ist keine beliebige Lücke, sondern die einzige Stelle, an der die KI-VO für zumindest genau diese beiden Hochrisiko-KI-Systeme (Anhang III Nr. 5 lit b und c KI-VO) überhaupt an das Recht aus Art. 22 Abs. 3 DSGVO anschließen könnte: Das Recht der betroffenen Person, das Eingreifen einer Person zu erwirken, den eigenen Standpunkt darzulegen und die Entscheidung anzufechten (EuGH, C-634/21). Ein Beschwerdemechanismus nach Art. 27 Abs. 1 lit. f KI-VO wäre der naheliegende operative Kanal, über den betroffene Personen dieses Recht bei einem Hochrisiko-KI-System praktisch geltend machen. Die KI-VO verlangt aber nur, dass ein solcher Mechanismus beschrieben wird — nicht, dass er diesen

[5] Anm.: Version 1.0, 7. Juli 2026, Konsultation bis Ende Oktober 2026 — der folgende Befund bezieht sich auf den Entwurf und steht unter Vorbehalt der finalen Fassung.

Rechten inhaltlich entspricht oder eine bloß formale Bestätigung ausschließt. Der Anschluss an Art. 22 Abs. 3 DSGVO wäre naheliegend gewesen; die Norm stellt ihn nicht her.

Damit trifft die Lücke aus Punkt 3.1 hier ein zweites Mal, in einer neuen Form: Wo Punkt 3.1 zeigte, dass formale Beteiligung keine Kontrolle ist, zeigt Art. 27 KI-VO, dass der Gesetzgeber diesen bereits etablierten Maßstab nicht einmal dort heranzieht, wo er ihn am direktesten hätte anwenden können.

Die zweite Lücke betrifft das Verhältnis zu Art. 14 KI-VO — und sie ist keine bloße Organisationsfrage. Art. 14 KI-VO verlangt laufende, interne Aufsicht während des Betriebs, ausgeübt von einer intern bestimmten Aufsichtsperson. Art. 27 KI-VO verlangt einen externen, reaktiven Kanal, der erst auf Initiative der betroffenen Person nach einer Entscheidung greift. Das sind strukturell zwei verschiedene Kontrollfunktionen — die eine soll Fehler erkennen, bevor sie wirken, die andere korrigiert im Einzelfall danach. Art. 27 KI-VO klärt nicht, ob und wie Erkenntnisse aus dem Beschwerdemechanismus in die laufende Aufsicht nach Art. 14 KI-VO zurückfließen, noch ob dieselbe Person beide Funktionen ausüben darf.

Fallen sie bei kleineren Betreibern aus Ressourcengründen zusammen, ist das keine bloße Doppelbelastung, sondern eine Schwächung der Aufsicht selbst: Beide Funktionen konkurrieren um dieselbe Zeit, dieselbe Aufmerksamkeit, dieselbe Person — und ohne Rückkopplung bleibt die einzelfallbezogene Erkenntnis aus dem Beschwerdemanagement von der systemischen Aufsicht getrennt, die sie eigentlich informieren müsste.

4. Der regulatorische Anspruch – Art 14 KI-VO

4.1 Ein Risiko, das die Norm selbst schon kennt

Art. 14 KI-VO verlangt, dass Hochrisiko-KI-Systeme so konzipiert sind, dass natürliche Personen sie während des Einsatzes wirksam beaufsichtigen können — inklusive geeigneter Mensch-Maschine-Schnittstelle. Die Aufsichtsperson soll die Ausgabe richtig interpretieren, sich gegen die Nutzung entscheiden und in den Betrieb eingreifen oder ihn per Stopptaste unterbrechen können.

Bemerkenswert daran: Die Norm benennt das eigentliche Risiko bereits selbst — Art. 14 Abs. 4 lit. b KI-VO verlangt wörtlich, dass Aufsichtspersonen sich der „möglichen Neigung zu automatischem oder übermäßigem Vertrauen" in die Systemausgabe „bewusst bleiben". Eine über den Normtext hinausgehende Erläuterung dazu fehlt bislang — eine Leitlinie nach Art. 96 KI-VO speziell zu Art. 14 KI-VO gibt es noch nicht. Automation Bias ist damit im Gesetz selbst benannt, aber nicht operationalisiert. Was fehlt, ist die Antwort auf die naheliegende Anschlussfrage: Wie soll eine Aufsichtsperson dieses Bewusstsein aufrechterhalten, wenn Zeitdruck, Durchsatzanreize und die schiere Menge an zu prüfenden Ergebnissen strukturell in die entgegengesetzte Richtung wirken?

4.2 Ein Phänomen, das sich nicht durch Bewusstsein beheben lässt

Automatisierungsbias muss nicht erst über Sekundärquellen eingeführt werden — die KI-VO definiert ihn selbst, in Art. 14 Abs. 4 lit. b KI-VO: Die Neigung zu automatischem oder übermäßigem Vertrauen in die Ausgabe eines Hochrisiko-KI-Systems. **Bemerkenswert ist die Rechtsfolge, die der Gesetzgeber daraus zieht: Eine reine Bewusstseinspflicht.** Die Aufsichtsperson soll sich der Neigung „bewusst bleiben" — eine kognitive Anforderung an den Menschen, keine Gestaltungspflicht des Anbieters, etwa durch Friktion vor Bestätigung, Konfidenzintervalle statt Punktwerte oder Zwangsverzögerung bei kritischen Entscheidungen.

Die kognitionspsychologische Forschung widerspricht diesem Ansatz: Achtsamkeitsappelle reduzieren Automatisierungsbias kaum, weil der Effekt strukturell und nicht willentlich steuerbar ist. Die Norm

begegnet damit einem Phänomen, das sich kaum durch Willenskraft kontrollieren lässt, mit einer reinen Willenskraft-Anforderung.

Dass Automation Bias real und robust ist, ist breit dokumentiert. Der International AI Safety Report 2026[6] hält fest, dass die Tendenz, automatisierten Ergebnissen unkritisch zu vertrauen und widersprechende Informationen zu ignorieren, eigenständige Kompetenz untergräbt, aktives Nachdenken entmutigt und eigenständiges Urteil schwächt. Bereits assistierte KI-Nutzung, etwa beim Schreiben mit einem Chatbot, verschiebt nachweislich die Meinungen der Nutzer in Richtung der Modellvorschläge.

Die rechtswissenschaftliche Literatur hat diesen Befund längst mit der KI-VO konfrontiert. Laux und Ruschemeier zeigen in „Automation Bias in the AI Act"[7], dass die Norm zwar Bewusstsein für Automation Bias verlangt, aber weder Design noch Nutzungskontext als eigentliche Ursache adressiert — und stellen die Frage, ob das Risiko nicht direkt reguliert werden müsste, statt es dem Bewusstsein Einzelner zu überlassen. Samir Passi kommt in „Agentic AI has a Human Oversight Problem"[8] zu einem verwandten Ergebnis: Die Annahme, ein Mensch könne autonome Agenten wirksam beaufsichtigen, unterschätzt systematisch, wie stark die Interaktionsform selbst die Prüffähigkeit untergräbt.

4.3 Konkretisierung in der Warteschleife

Die harmonisierten Normen prEN 18229-1 bzw. prEN 18229-3 („AI Trustworthiness Framework — Part 1: Logging; Part 3: transparency and human oversight") sollen strukturierte Umsetzungsmethoden für Art. 14 KI-VO liefern. Sie befinden sich noch in Entwicklung. Selbst wenn sie verfügbar wären, änderte das nichts am eigentlichen Befund: Eine technische Norm kann fehlende Legal Governance nicht ersetzen. EN 18229-1 bzw. -3 liefert bestenfalls Umsetzungsmethoden für das, was Art. 14 KI-VO bereits verlangt — sie kann keine Verifikationspflicht schaffen, wo die Norm selbst keine vorsieht (4.6), keinen Wirksamkeitsstandard definieren, wo der Gesetzestext nur Umsetzbarkeit fordert, und keine Kompetenzprüfung einführen, wo Art. 4 KI-VO und Art. 26 Abs. 2 KI-VO keine verlangen. Eine harmonisierte Norm konkretisiert eine bestehende Rechtspflicht, sie erzeugt keine neue. Bis eine Konkretisierung vorliegt — und selbst danach — bleibt die eigentliche Lücke: Eine Norm, deren zentrales Risiko die Kommission benennt, aber nicht operationalisiert, und die dieses Versäumnis nicht an eine technische Norm delegieren kann.

4.4 Rohdaten sind keine Baseline

Art. 12 KI-VO verpflichtet Anbieter, Hochrisiko-KI-Systeme so auszulegen, dass eine automatische Protokollierung von Ereignissen möglich ist — zur Ermittlung von Risikosituationen (Art 12 Abs. 2 lit. a KI-VO) und zur Betriebsüberwachung durch den Betreiber nach Art. 26 Abs. 5 KI-VO (Art 12 Abs. 2 lit. c KI-VO). Damit entsteht eine Aufteilung, die die Norm nicht ausdrücklich benennt: Der Anbieter baut die Protokollierungsfähigkeit ein, der Betreiber überwacht und bewahrt auf (Art. 26 Abs. 6 KI-VO). Keiner der beiden Artikel verpflichtet jedoch zur aktiven, kontinuierlichen Auswertung dieser Protokolle — die Aufteilung regelt, wer protokolliert und aufbewahrt, nicht, wer auswertet.

Das ist mehr als ein organisatorisches Detail: Es trifft direkt die Anforderung aus Art. 14 Abs. 4 lit. a KI-VO, wonach die Aufsichtsperson den Betrieb überwachen können muss, um Anomalien, Fehlfunktionen und unerwartete Leistung zu erkennen — und genau dieses Erkennen setzt notwendig einen Referenzpunkt voraus, sonst gibt es kein „unerwartet". Genau diesen Referenzpunkt liefert Art. 12 KI-VO nicht: Er verlangt die Aufzeichnung diskreter Ereignisse, aber keine Baseline-Metriken bei

[6] Herausgeber britische Regierung; Februar 2026; https://arxiv.org/pdf/2602.21012.

[7] Laux, J., Ruschemeier, H.; Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI; Februar 2025; https://arxiv.org/abs/2502.10036.

[8] Passi, S.; Agentic AI has a Human Oversight Problem; Oktober 2025; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5529058.

Inverkehrbringen, keine kontinuierliche aggregierte Auswertung, keine definierten Abweichungsschwellen. Ein graduelles Driften der Ausgabeverteilung lässt sich ohne Baseline und Vergleichsdaten auch im Nachhinein nicht aus Einzelprotokollen rekonstruieren — und Art. 12 KI-VO schreibt inhaltlich kaum vor, welche Daten überhaupt protokolliert werden müssen, sodass selbst die Rohdaten für eine solche Rekonstruktion nicht gesichert sind. Als eigenständig zu erkennendes Phänomen adressiert Art. 12 KI-VO Drift damit an keiner Stelle.

Damit trifft die Aufsichtsperson eine zweite, von Automation Bias (4.2) unabhängige Lücke: Selbst wer Automation Bias erfolgreich widersteht, bekommt von der Protokollierungsnorm nicht die Grundlage, die Art. 14 KI-VO voraussetzt. Die Kontrollinstanz ist damit nicht nur psychologisch geschwächt, sondern es fehlt ihr bereits die technische Infrastruktur, auf der wirksame Aufsicht überhaupt aufbauen könnte.

Dass diese Lücke nicht durch andere Regelwerke geschlossen wird, bestätigt ein Randbefund an anderer Stelle der KI-VO selbst: Art. 72 Abs. 1 UAbs. 3 KI-VO ordnet für Anhang-III-Nr.-5-Systeme von Finanzinstituten an, dass die Post-Market-Monitoring-Pflicht zusätzlich gilt („gilt auch für") — anders als bei Art. 17 Abs. 4, Art. 18 Abs. 3 und Art. 26 Abs. 5 UAbs. 2 KI-VO, wo DORA-Konformität als Erfüllung der KI-VO-Pflicht fingiert wird („gilt als erfüllt"). Hätte der Gesetzgeber angenommen, DORA-Governance decke auch das Beobachten von Modellverhalten und Drift ab, hätte es für Art. 72 KI-VO keinen Grund für eine Zusatzpflicht statt einer weiteren Fiktion gegeben. DORA (vgl. etwa Art. 5 f., 28 DORA) enthält tatsächlich keine Verpflichtung zur systematischen Beobachtung von KI-Modellverhalten über den Lebenszyklus. Der Gesetzgeber widerlegt an dieser einen Stelle also selbst die Annahme, auf der die Fiktionen an den drei Nachbarstellen beruhen, ohne sie dort zu korrigieren — ein Beleg dafür, dass die fehlende Baseline aus Art. 12 KI-VO keine übersehene Randfrage ist, sondern eine Lücke, die der Gesetzgeber an anderer Stelle unbeabsichtigt selbst offenlegt.

Für Finanzinstitute im DORA Anwendungsbereich, die ein Hochrisiko-KI-System nach Anhang III Nr. 5 KI-VO einsetzen, heißt das konkret: Ihr bestehendes ICT-Risikomanagement nach DORA deckt die Beobachtung von Modellverhalten und Drift nicht ab. Art. 72 KI-VO verlangt diese Beobachtung jedoch zusätzlich und ausdrücklich, ohne dass ein Rückgriff auf eine DORA-Fiktion sie ersetzen könnte. Ein Institut, das sein KI-Monitoring auf die bestehende DORA-Governance stützt, erfüllt damit weder die stille Erwartung des Gesetzgebers noch die explizite Zusatzpflicht aus Art. 72 KI-VO — es braucht eine eigenständige, modellspezifische Beobachtungsfähigkeit, die weder DORA noch Art. 12 KI-VO liefert und die es selbst aufbauen muss.

4.5 Verständlich – aber für wen?

Art. 13 Abs. 2 KI-VO verlangt Betriebsanleitungen in „präziser, vollständiger, korrekter und eindeutiger" sowie „verständlicher" Form, ohne den Adressaten dieser Verständlichkeit zu benennen. Das Verbraucherschutzrecht verfügt mit dem durchschnittlich informierten, aufmerksamen und verständigen Durchschnittsverbraucher (st. Rspr. seit Gut Springenheide, C-210/96) über einen ausgearbeiteten Referenzmaßstab für genau diese Frage. Art. 13 KI-VO kennt keinen vergleichbaren Maßstab — ob „verständlich" an einem IT-Compliance-Profi, an einer Person ohne technischen Hintergrund oder an einer fiktiven Referenzfigur wie der Sorgfalt eines ordentlichen Kaufmanns zu messen ist, bleibt offen.

Die Analogie zur DSGVO liegt nahe, ist aber nicht verankert. Art. 12 Abs. 1 DSGVO verlangt „präzise, transparente, verständliche und leicht zugängliche" Form „in klarer und einfacher Sprache" — nahezu wortgleich. Dort existieren mit den WP260-Leitlinien der ehemaligen Art.-29-Gruppe (heute EDSA) tatsächlich Konkretisierungskriterien: Zielgruppenorientierung, Layered Notices, Lesbarkeitstests. Die KI-VO verweist trotz des nahezu identischen Wortlauts nicht darauf.

Für Art. 14 KI-VO ist das unmittelbar relevant: Die Betriebsanleitung bildet die Grundlage, auf der die Aufsichtsperson lernen soll, wie sie das System bedient und wo seine Grenzen liegen (Art. 14 Abs. 4 lit. a KI-VO). Ist sie für die tatsächliche Zielgruppe nicht verständlich, ist die Kontrollfähigkeit bereits an der Quelle geschwächt — noch bevor Automation Bias oder die fehlende Baseline aus 4.4 überhaupt zum Tragen kommen.

4.6 Fünf Fähigkeiten, keine Verifikation

Art. 14 Abs. 3 KI-VO lässt Anbietern die Wahl zwischen zwei gleichrangigen Alternativen: Vorkehrungen, die vor Inverkehrbringen ins System eingebaut werden (lit. a), oder solche, die nur „geeignet sind, vom Betreiber umgesetzt zu werden" (lit. b). Eine Kaskadenlogik, die der technisch robusteren Einbaulösung Vorrang einräumt, wenn sie machbar ist, sieht die Norm nicht vor. Problematisch ist vor allem der Maßstab von lit. b: „Geeignet, umgesetzt zu werden" beschreibt Umsetzbarkeit, nicht Wirksamkeit. Der Anbieter erfüllt seine Pflicht bereits durch die Bestimmung einer „geeigneten" Vorkehrung — ob sie beim konkreten Betreiber tatsächlich wirkt, hängt von dessen Kompetenz und Kontext ab, die der Anbieter beim Design nur abstrakt kennen kann. Es ist derselbe Erwartungs- statt Tatsachenmaßstab wie in Art. 9 Abs. 5 KI-VO, der die Risikominderung an Kenntnissen ausrichtet, „die vom Betreiber erwartet werden können" — wer diese Erwartung festlegt, sagt die Norm nicht. In der Praxis ist es der Anbieter selbst, der beim Design eine Annahme über die Kompetenz eines ihm unbekannten künftigen Betreibers trifft, ohne dass diese Annahme irgendwo verifiziert oder auch nur dokumentiert werden müsste. Der Maßstab ist damit nicht nur erwartungs- statt tatsachenbasiert, sondern wird von der Partei bestimmt, die am wenigsten über den tatsächlichen Betreiber weiß.

Dasselbe Muster wiederholt sich in Art. 14 Abs. 4 KI-VO: Fünf Buchstaben verlangen fünf Fähigkeiten — verstehen, bewusst bleiben, interpretieren, beschließen, eingreifen —, und keine davon ist mit einer Verifikationspflicht verbunden. Es gibt keine Eignungsprüfung, keine Simulation unter Zeitdruck, keine Messung der Interventionsrate, kein Kriterium dafür, wann eine Aufsichtsperson „angemessen und verhältnismäßig in der Lage" ist. Art. 14 Abs. 4 lit. c KI-VO verschärft das durch einen Zirkelbezug zu Art. 13 KI-VO: Die Interpretationsfähigkeit ist an „vorhandene Interpretationsinstrumente" gekoppelt — liefern diese, wie in 4.5 gezeigt, keine verlässliche Erklärung, sinkt der Maßstab mit ihnen.

Besonders deutlich wird die Lücke bei Art. 14 Abs. 4 lit. e KI-VO, der Eingriffs- und Stoppfähigkeit verlangt. Wirksamer Eingriff setzt voraus, nicht nur zu erkennen, dass er nötig ist, sondern auch, in welchem Zeitfenster er noch wirkt. Art. 14 Abs. 3 KI-VO nennt dafür lediglich den „Grad der Autonomie" als Angemessenheitskriterium — ohne Reaktionszeitfenster, ohne Pflicht zur Synchronisation mit der Verarbeitungsgeschwindigkeit des Systems. Eine Stopptaste, die erst nach Ablauf des relevanten Zeitfensters betätigt werden kann, erfüllt Art. 14 Abs. 4 lit. e KI-VO damit formal, ohne real zu wirken — genau die Lücke, die OWASP ASI10 (Kapitel 6) mit der Trennung „Preview"/„Execute" und risikoadaptiver Aufsicht informell zu schließen versucht, ohne dass die Norm eine entsprechende Pflicht kennt.

Offen bleibt zudem, was nach Betätigung dieser — in der Praxis oft so genannten — Kill-Switch-Funktion geschehen soll: Der Wortlaut verlangt nur einen „sicheren Zustand", ohne ihn zu definieren. Die naheliegende Analogie zur Zug-Notbremse trägt hier nicht: Sie beschreibt einen einzelnen, klar abgrenzbaren Bewegungszustand. Ein agentisches System dagegen hat zum Zeitpunkt des Eingriffs möglicherweise bereits Transaktionen initiiert, E-Mails versendet oder andere Systeme aufgerufen. Ob „sicherer Zustand" in diesem Fall nur den weiteren Prozess stoppt oder bereits ausgelöste Aktionen zurückrollt, lässt die Norm offen — weder eine Rollback- noch eine Verifikationspflicht ist vorgesehen. Bei verteilten Software-Systemen mit irreversiblen Nebenwirkungen bleibt „sicherer Zustand" damit ein unspezifizierter Begriff — ein Punkt, den viele Organisationen, die gerade Hochrisiko-KI-Systeme

beschaffen, übersehen: Die Frage, was nach einem Stopp mit bereits ausgelösten Aktionen passiert, gehört in die Beschaffungsprüfung, nicht erst in den Ernstfall.

4.7 Kompetenz – nur vermutet

Die Fähigkeiten aus Art. 14 Abs. 4 KI-VO setzen Kompetenz voraus. Die einzige Norm, die dafür bislang eine korrespondierende Pflicht vorsah, war Art. 4 KI-VO: Ein „ausreichendes Maß an KI-Kompetenz" sicherzustellen — ein Ergebnisstandard. Mit dem KI-Omnibus, dem der Rat am 29. Juni 2026 zugestimmt hat (Veröffentlichung im Amtsblatt stand zum Zeitpunkt der Veröffentlichung dieses Whitepapers noch aus), wird daraus eine bloße Förderpflicht: Kompetenz fördern, ohne ein Niveau zu garantieren. Gleichzeitig verschieben sich die Anwendungsfristen für Hochrisiko-Systeme — Anhang III-KI-VO-Systeme auf den 2. Dezember 2027, Anhang I-KI-VO-Systeme auf den 2. August 2028. Damit fällt das letzte Element weg, das direkt bei der Kompetenz des Menschen ansetzte.

Auf Betreiberseite zeigt sich dasselbe Muster: Art. 26 Abs. 2 KI-VO überträgt die Aufsicht Personen mit „erforderlicher Kompetenz, Ausbildung und Befugnis" — ohne Eignungsprüfung, ohne Mindestqualifikation, ohne Auffrischungspflicht. Wer beurteilt, ob die Kompetenz ausreicht, bleibt offen. Der in der Praxis verbreitete „AI Officer"-Titel ändert daran nichts, da die KI-VO diese Rolle nicht kennt — er ist eine Marktkonstruktion, die die Lücke besetzt, ohne einen Maßstab zu liefern, ob die Person die Anforderungen aus Art. 14 Abs. 4 KI-VO tatsächlich erfüllt. Ein bestellter AI Officer ist damit nicht automatisch eine befähigte Person im Sinne der Norm — und selbst ein exzellenter AI Officer, der ein tadelloses KI-Managementsystem nach ISO 42001 aufsetzt, ist damit noch keine taugliche Aufsichtsperson für ein konkretes System: Die Kompetenz nach Art. 14 Abs. 4 KI-VO ist tool- und risikobezogen, nicht organisatorisch. Governance-Kompetenz auf Systemebene und Beaufsichtigungskompetenz für ein einzelnes Kreditvergabesystem sind zwei verschiedene Fähigkeiten, die Art. 26 Abs. 2 KI-VO unter demselben unbestimmten Begriff zusammenfasst.

4.8 Genau genug, um zu genügen – nicht mehr

Art. 15 Abs. 1 KI-VO verlangt ein „angemessenes Maß" an Genauigkeit, Robustheit und Cybersicherheit, ohne diesen Maßstab zu konkretisieren — und ohne den Begriff „Cybersicherheit" selbst zu definieren oder auf eine bestehende Definition zu verweisen. Dass eine solche Definition verfügbar wäre, zeigt Art. 6 Nr. 3 NIS-2-Richtlinie, der "Cybersicherheit" ausdrücklich über einen Verweis auf Art. 2 Nr. 1 der Cybersecurity Act-Verordnung (EU) 2019/881 bestimmt. Die KI-VO hätte diesen Verweis übernehmen können — sie tut es nicht, sondern lässt den zentralen Begriff des Artikels unverankert.

Art. 15 Abs. 2 KI-VO verschärft die Unschärfe noch, statt sie aufzulösen: Die Kommission fördert Benchmarks ausdrücklich nur für Genauigkeit und Robustheit — Cybersicherheit, im Artikeltitel prominent genannt, fehlt auch hier in der Aufzählung. Art. 15 Abs. 3 KI-VO vertieft die Asymmetrie weiter: In die Betriebsanleitung müssen nur Genauigkeitsmetriken, keine Robustheitswerte — von Cybersicherheitswerten ganz zu schweigen.

Das ist für Art. 14 KI-VO direkt relevant. Art. 14 Abs. 4 lit. a KI-VO verlangt das Erkennen von „Anomalien, Fehlfunktionen und unerwarteter Leistung" — eine Fähigkeit, die belastbare Referenzwerte voraussetzt. Bleiben diese Werte vage und fehlen Robustheitsdaten in der Betriebsanleitung, fehlt der Aufsichtsperson ein weiteres Stück der Grundlage, die sie für diese Aufgabe bräuchte. Es ist dasselbe Muster wie die fehlende Baseline aus Art. 12 KI-VO (4.4) — diesmal nicht bei den Protokolldaten, sondern bei den Leistungskennzahlen selbst, auf denen jede Anomalieerkennung aufbauen müsste.

Für NIS2- und DORA-Verantwortliche ist diese Lücke in der Praxis besonders unbequem: Welchen Maßstab legt man an ein KI-System an, wenn die KI-VO selbst keinen liefert? Der naheliegende Rückgriff auf den CRA trägt nur dort, wo der CRA überhaupt greift — bei reinen SaaS-/API-Agentic-AI-Systemen

und bei KI-Komponenten in bereits sektorregulierten Produkten ist das gerade nicht der Fall. Und selbst wo der CRA anwendbar ist, hat er KI-spezifische Logik nicht mitgedacht, agentische Systeme schon gar nicht (vgl. Kapitel 5). Damit bleibt eine Grundfrage offen, die weit über Compliance hinausgeht: Wie soll ein Hochrisiko-KI-System sicher betrieben werden, wenn Anomalien nicht einmal sinnvoll erkennbar sind — jedes durchschnittliche SOC ist an dieser Stelle besser ausgestattet als der gesetzliche Rahmen, der das System eigentlich absichern soll.

4.9 Verantwortlich, aber nicht handlungsfähig

Art. 26 Abs. 4 KI-VO verpflichtet Betreiber, „soweit die Eingabedaten ihrer Kontrolle unterliegen", für Zweckentsprechung und Repräsentativität zu sorgen — zwei Begriffe, die die Norm selbst nicht definiert. Beides sind keine Fragen, die sich intuitiv beantworten lassen: Ob eine Datengrundlage tatsächlich der Zweckbestimmung entspricht und statistisch repräsentativ ist, lässt sich seriös nur mit einschlägiger fachlicher — insbesondere statistischer — Expertise beurteilen, die beim Betreiber nicht vorausgesetzt werden kann. Dasselbe Kompetenzproblem wie bei Art. 4 und Art. 26 Abs. 2 KI-VO (4.7) taucht hier ein zweites Mal auf, diesmal nicht bei der Aufsichtsperson, sondern bei der Datenbewertung selbst.

Hinzu kommt ein struktureller Zeitpunkt-Fehler: Repräsentativität ist keine Eigenschaft, die sich einmalig bei Inbetriebnahme feststellen und dann als erledigt betrachten lässt — sie kann über die Laufzeit des Systems driften, wenn sich die Population der tatsächlichen Eingabedaten verändert.

Um das nachzuweisen oder auch nur zu widerlegen, müsste der Betreiber die Eingabedaten über die gesamte Betriebsdauer aufbewahren und regelmäßig neu bewerten. Art. 26 KI-VO verlangt das nicht: Art 26 Abs. 6 KI-VO schreibt eine Aufbewahrungspflicht nur für die vom System automatisch erzeugten Protokolle vor (mindestens sechs Monate), nicht für die Eingabedaten selbst. Dieselbe Lücke wie bei der fehlenden Baseline aus Art. 12 KI-VO (4.4) — Repräsentativität wird als Pflicht formuliert, aber ohne die Dateninfrastruktur, die nötig wäre, um sie über die Zeit überhaupt zu überprüfen.

5. Von der Compliance zur Meldepflicht - CRA

Der Cyber Resilience Act (Verordnung (EU) 2024/2847) verlagert die Frage von der Compliance- auf die Produktsicherheitsebene. Fällt ein agentisches KI-System unter den CRA-Produktbegriff — „Produkt mit digitalen Elementen“ nach Art. 3 Z 1 CRA —, greifen die Anforderungen aus Anhang I CRA: Security-by-Design-and-Default (Teil I), kontinuierliches Schwachstellenmanagement über den Produktlebenszyklus inklusive aktiver Beobachtungspflicht (Teil II Nr. 3).

Trust Exploitation im Sinne von OWASP ASI09 ist damit nicht länger nur eine abstrakte Sicherheitskategorie, sondern potenziell eine unter Art. 14 CRA meldepflichtige Schwachstellenklasse — meldepflichtig ist dabei der Hersteller (Art. 3 Nr. 13 CRA), nicht der Betreiber. Wird sie aktiv ausgenutzt, greift ein gestaffeltes Fristenregime: Frühwarnung binnen 24 Stunden, ausführlichere Meldung binnen 72 Stunden, Abschlussbericht binnen 14 Tagen nach Verfügbarkeit einer Abhilfemaßnahme — jeweils gleichzeitig an das zuständige CSIRT und die ENISA. Zusätzlich verpflichtet Art. 14 Abs. 8 CRA den Hersteller zur aktiven Benachrichtigung seiner Nutzer über die Schwachstelle und verfügbare Updates — eine Pflicht, die strukturell voraussetzt, dass der Hersteller weiß, wer seine Nutzer sind und was sie zur Einordnung brauchen. Fehlt, wie in 4.5 gezeigt, bereits die verständliche Betriebsanleitung nach Art. 13 KI-VO, ist diese Benachrichtigungspflicht faktisch nicht erfüllbar, obwohl sie rechtlich unbedingt gilt.

Das verschiebt die rechtliche Logik erheblich. Aus der Frage, ob ein Anbieter Art. 14 KI-VO „im Rahmen des Möglichen" eingehalten hat, wird die Frage, ob ein Hersteller eine bekannte, klassifizierte Schwachstellenkategorie im Schwachstellenmanagement berücksichtigt und fristgerecht gemeldet hat.

Das ist schärfer und besser justiziabel: Bußgelder bis zu 15 Millionen Euro oder 2,5 % des weltweiten Jahresumsatzes machen aus einer regulatorischen Erwartung eine unmittelbar sanktionsbewehrte Pflicht.

Bemerkenswert am Rande: Der CRA-Herstellerbegriff nutzt dieselbe Auslösefigur wie Art. 25 KI-VO für den Anbieterstatus — eigener Name/eigene Marke oder Durchführen einer wesentlichen Änderung, die über ein harmloses Sicherheitsupdate hinausgeht. Naheliegend, aber bislang weder von der Kommission noch in der CRA-Guidance ausdrücklich bestätigt: Wer durch eine solche Änderung vom Betreiber zum Anbieter nach KI-VO mutiert, dürfte damit vermutlich zugleich den Hersteller-Tatbestand nach CRA für dasselbe Produkt erfüllen — eine Einschätzung, die sich aus der Parallelität der beiden Auslösefiguren ergibt, aber derzeit keine amtliche Bestätigung hat. Für Betreiber, die ein Hochrisiko-KI-System unter eigenem Namen oder eigener Marke in Verkehr bringen — was in der Praxis nicht selten vorkommt —, heißt das: Sie kaufen sich damit ein zweites, eigenständiges Pflichtenregime ein, das auf dem Radar oft gar nicht vorkommt, weil der Fokus meist ausschließlich auf der KI-VO-Anbieter-/Betreiberpflicht liegt.

6. Die Lücke ist nicht hypothetisch — OWASP ASI09 und ASI10

Dass die in Kapitel 3 bis 5 hergeleitete regulatorische Lücke real ausnutzbar ist und nicht nur theoretisch besteht, bestätigt das OWASP Top 10 for Agentic Applications 2026 (ASI01–ASI10, Dezember 2025, entwickelt mit über 100 Sicherheitsexpertinnen und -experten, unter anderem unterstützt von NIST, Microsoft, NVIDIA).

ASI09 — Human-Agent Trust Exploitation beschreibt gezielt die Manipulation menschlichen Vertrauens in Agenten-Outputs als eigenständigen Angriffsvektor: Angreifer gestalten Ausgaben so, dass sie Vertrauen erzeugen und inhaltliche Prüfung gezielt umgehen. Das ist keine Fahrlässigkeit der Aufsichtsperson, sondern eine Technik, die gezielt die Automation-Bias-Anfälligkeit aus Kapitel 4 adressiert und ausnutzt.

ASI10 — Rogue Agents beschreibt das Extremszenario: Ein Agent, der vom beabsichtigten Verhalten abweicht und mit schädlicher Autonomie handelt — autorisiert, vertraut, fehlausgerichtet, schwer zu erkennen. Als Gegenmaßnahmen nennt OWASP adaptive Aufsicht mit steigendem Risiko, eine klare Trennung zwischen „Preview" und „Execute", unveränderliche Audit-Logs und — ausdrücklich — Training gegen Automation Bias und emotionales Vertrauen.

Der Kreis schließt sich damit: Was die DSGVO-Rechtsprechung als Scheinkontrolle beschreibt, was Art. 14 KI-VO als Risiko benennt, ohne es zu operationalisieren, und was der CRA als potenziell meldepflichtige Schwachstelle einordnet, ist in der Fachwelt längst als eigenständige, aktiv ausnutzbare Angriffskategorie klassifiziert. Die Lücke ist nicht hypothetisch.

7. Kontrolle ohne Selbstschutz trägt nicht

Rosenthals Unterscheidung zwischen „Human in the Loop" und „Human controls the loop" ist richtig — und unvollständig, wenn isoliert betrachtet. „Controls the loop" trifft zu: Bloße nachträgliche Prüfung reicht nicht. Aber auch echte Kontrolle nützt wenig, wenn der Gesetzgeber sie als Schutzmaßnahme behandelt, nicht als Ziel: Kontrollfähigkeit, die nicht selbst gegen Angriffe abgesichert ist, bleibt kompromittierbar. Genau das ist in keinem der drei Rechtsregime der Fall — und genau das macht die Kontrollinstanz zum Angriffsziel.

Automation Bias ist der Mechanismus, über den formale Kontrolle unbemerkt zu Scheinkontrolle wird. ASI09 zeigt, dass dieser Mechanismus gezielt angreifbar ist. Die DSGVO-Rechtsprechung zeigt, dass Gerichte formale Kontrolle bereits Jahre zuvor als unzureichend zurückgewiesen haben. Art. 14 KI-VO

benennt das Risiko, ohne es zu schließen. Der CRA verschiebt die Konsequenzen auf die Produktebene, löst aber die Governance-Lücke dahinter nicht auf.

Art. 14 Abs. 2 KI-VO macht das im Normtext selbst greifbar: Menschliche Aufsicht dient der Minimierung von Risiken, die auch bei voller Compliance mit den übrigen Anforderungen fortbestehen. Gestaffelte Sicherungsebenen, jede als Auffangnetz für die vorherige, sind für sich genommen eine legitime, verbreitete Regulierungstechnik — kein Eingeständnis von Schwäche. Die Absicht ist ein funktionierender Backstop. Gegengelesen mit den Lücken aus Kapitel 4 — fehlende Baseline (Art. 12), unbestimmte Verständlichkeit (Art. 13), Kaskadenlogik- und Verifikationslücken (Art. 14 Abs. 3 und 4), generische Genauigkeits- und Robustheitsangaben (Art. 15), gelockerte Kompetenzanforderungen (Art. 4, Art. 26 Abs. 2) — legt Art. 14 Abs. 2 unfreiwillig offen: Dieser Backstop durchläuft keine eigene Tragfähigkeitsprüfung. Das ist nicht die Intention des Normtexts, sondern seine Konsequenz.

Rosenthals Beobachtung aus einem einzelnen Vortrag ist damit Symptom eines strukturellen Problems, das sich durch das gesamte europäische Digitalregulierungsrecht zieht: Menschliche Kontrolle wird als Lösung vorgeschrieben, ohne ihre eigene Angreifbarkeit mitzudenken.

8. Compliance ist nicht Sicherheit

8.1 Conclusio

Der Gesetzgeber schreibt in DSGVO, KI-VO und CRA dieselbe Schwachstelle fest: Menschliche Aufsicht gilt in allen drei Regimen als Schutzmaßnahme, nicht als eigenständig abzusicherndes Ziel — deshalb bleibt sie angreifbar. Solange keines der drei Regime das adressiert, bleibt „wirksame menschliche Aufsicht" eine rechtliche Fiktion, die sich durch Automation Bias und gezielte Trust Exploitation gleichermaßen unterlaufen lässt. Für Anbieter und Betreiber heißt das konkret: Compliance mit dem Normwortlaut erzeugt keine Sicherheit, sondern genau die vorhersehbare Angriffsfläche, die dieses Whitepaper beschrieben hat.

8.2 Praxisempfehlungen

- **Die Kontrollinstanz selbst red-teamen, nicht nur den Agenten.** Wer testet, ob der Agent manipulierbar ist, aber nie testet, ob der Mensch am Kontrollpunkt manipulierbar ist, hat die Hälfte der Angriffsfläche ignoriert. Trust-Exploitation-Szenarien gehören ins Pentesting-Scope.
- **Aufsicht, die nie widerspricht, ist keine Aufsicht.** Liegt die Freigabequote nahe 100 %, kontrolliert niemand mehr — das Gremium nickt ab. Das gehört als Kennzahl ins Monitoring, nicht als Nebensächlichkeit.
- **Erklärbarkeit vor Ausführung, nicht danach.** Output-Prüfung nach dem Handeln des Agenten ist Schadensbegrenzung, keine Aufsicht. „Wirksame" Aufsicht im Sinne von Art. 14 KI-VO kann nur heißen: Eingriff vor irreversiblen Schritten möglich — sonst ist der Begriff inhaltsleer.
- **Automation Bias als Haftungsfrage behandeln, nicht als Trainingsfolie.** Wer „Human Oversight" auf dem Papier hat, aber nie geprüft hat, ob die Aufsichtsperson strukturell in der Lage ist, zu widersprechen (Zeitdruck, fehlende Fachkenntnis, Durchsatzanreize), hat die Pflicht formal, nicht faktisch erfüllt.
- **Keine Kontrolle ohne Exit.** Jede Aufsichtsfunktion braucht eine echte Stopp-Option mit Konsequenz — sonst ist „controls the loop" nur ein anderes Wort für „sieht zu".
- **Systembestandsaufnahme entlang aller drei Regime.** Welche Systeme unterliegen Art. 22 DSGVO, welche Art. 14 KI-VO, welche zusätzlich dem CRA-Produktbegriff? Die Überschneidung ist

der Bereich mit dem höchsten Risiko und dem größten Effizienzgewinn bei integrierter Betrachtung.

- **Kompetenz nachweisen statt vermuten.** Standardisierte Massenschulungen ohne Bezug zum konkreten System sind kein Nachweis von KI-Kompetenz nach Art. 14 Abs. 4 KI-VO — sie sind der Anschein davon. Ein dokumentierter, tool- und risikobezogener Education Plan, der die Aufsichtsperson konkret für das System qualifiziert, das sie tatsächlich beaufsichtigt, ist keine Kür, sondern die Voraussetzung dafür, dass die Kompetenzvermutung aus Art. 26 Abs. 2 KI-VO überhaupt trägt.
- **Anbieter-Nachweise aktiv einfordern, nicht nur entgegennehmen**. Betriebsanleitung, Baseline-Metriken, Robustheits- und Cybersicherheitsdaten, SBOM (Software Bill of Materials) — wo der Anbieter diese Nachweise nicht von sich aus liefert, muss der Betreiber sie vertraglich einfordern, bevor das System in Betrieb geht. Ebenso wichtig: Vertragsklauseln prüfen und zurückweisen, die Verantwortung für fehlende Anbieter-Nachweise pauschal auf den Betreiber verlagern — wer die Baseline nicht liefern kann, kann die Verantwortung dafür auch nicht vertraglich abwälzen.