

VISCHER

Menschliche Aufsicht über KI – funktioniert sie wirklich?

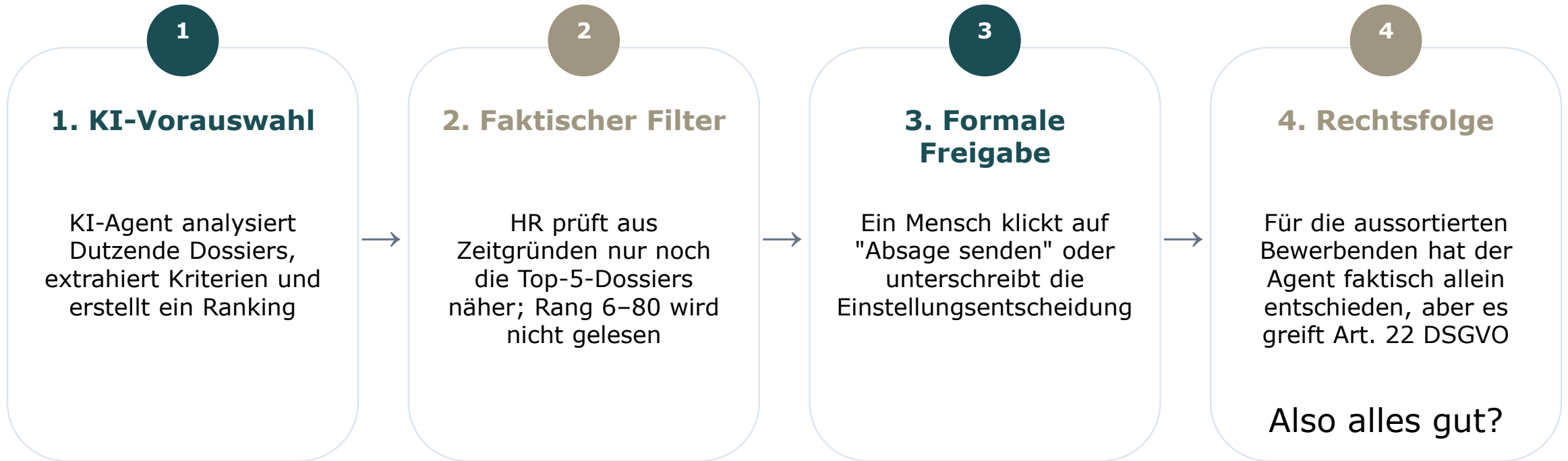
Zwischen Kontrollillusion und
echter Verantwortung

David Rosenthal
Partner

27. Oktober 2026 München



Ein Beispiel aus dem Recruiting ...



Die doppelte Illusion bei Art. 22 DSGVO

- **Illusion 1:** "Wir fallen gar nicht unter Art. 22 DSGVO"
- Art. 22 Abs. 1 DSGVO will verhindern, dass Menschen zum blossen Objekt algorithmischer Bewertungen werden
- Häufiges Verteidigungsargument in der Praxis: "Am Schluss schaut immer noch ein Sachbearbeiter drauf."
- *Fehlannahme:* Eine formale Zuständigkeit ohne reale Prüfzeit, ohne vollständige Aktenkenntnis und ohne gelebte Abweichungspraxis hebt die Automatisierung nicht wirklich auf
- Vgl. Schufa-Entscheid
- **Illusion 2:** "Der Mensch am Ende heilt das Risiko (Abs. 3)"
- Sind automatisierte Einzelentscheide trotz allem erlaubt, gewährt Art. 22 Abs. 3 DSGVO ein "Recht auf menschliches Gehör"
- Art. 14 Abs. 4 lit. b AI Act warnt vor einer "Automatisierungsverzerrung" (Automation Bias).
- *Paradoxon:* Das Gesetz versucht ein psychologisch-technisches Problem mit genau jener menschlichen Endkontrolle zu heilen, die in der Praxis systematisch überfordert ist

Lawsuit claims discrimination by Workday's hiring tech prevented people over 40 from getting hired



By Clare Duffy, CNN
5 min read · Updated 9:09 AM EDT, Thu May 22, 2025

Quelle: <https://edition.cnn.com/2025/05/22/tech/workday-ai-hiring-discrimination-lawsuit>

These: "Human in the Loop ist das neue Cookie-Banner"

Das klassische Cookie-Banner

1. Gedacht zur Verwirklichung einer informierten Selbstbestimmung
2. In der Realität: Reflexartiges Wegklicken innert Sekundenbruchteilen
3. Primär eine "Tick-Box"-Compliance-Vorgabe statt echtem Persönlichkeitsschutz
4. Niemand geht Hunderte von Partner-Infos- und Einstellungen auf einer Site tatsächlich durch

Illusion der Einwilligung

"Human in the Loop" bei KI

1. Standard-TOM in der DSFA: "Ergebnisse werden durch geschulte Mitarbeitende geprüft"
2. In der Realität: Übernahme eloquenter KI-Entwürfe/Entscheide unter Zeitdruck
3. Dient als juristisches Feigenblatt für Art. 22 DSGVO und Art. 14 AI Act
4. Verlagert das Systemrisiko auf das schwächste Glied der Kette

Illusion der Aufsicht

Warum der Mensch als Prüfinstanz ausfällt: Psychologie des Prüfers

Kognition Automation Bias & Druck

Denken ist anstrengend: Unter Zeitdruck übernehmen Menschen selbst widersprüchliche Maschinenurteile

Sprache Subtiles Framing

LLMs liefern flüssige, scheinbar objektive Begründungen ohne menschliche Stimmungsschwankungen

Kompetenz De-Skilling & No-Skilling

Wer nie selbst auf weissem Papier entwirft, verliert die Urteilskraft für das Review ("Lesen ist nicht Schreiben")

Routine Freigabe- Müdigkeit

Fragt ein Coding-Agent alle 60 Sekunden nach Freigabe, klicken Nutzer auf "Immer erlauben", wo doch fast alles richtig ist?

Hinzu kommt: Wir delegieren immer mehr Entscheide an die KI, so dass es illusorisch ist oder sein wird, dass alle wichtigen Entscheide vorher von einem Menschen geprüft werden ...

Die Illusion der menschlichen KI-Kontrolle: Wo "Human-in-the-Loop" heute scheitert

- **Latenz: Wenn die Zeit fehlt (Millisekunden)**

- Hochfrequenzhandel: Transaktionen und Risikosteuerung in Mikrosekunden
- Cyber-Abwehr: Echtzeit-Isolierung angegriffener Server vor der Verschlüsselung
- Infrastruktur: Autonome Kollisionsvermeidung und Stabilisierung von Stromnetzen

- **Volumen: Wenn die Masse erschlägt (Skalierung)**

- Finanzen & Kredite: Vollautomatische Betrugserkennung (Fraud Detection) und Kreditvergabe
- HR-Screening: Ungeprüfte "Schatten-Ablehnungen" in der Bewerber-Vorauswahl
- Informationsfluss: Millionenfach blockierte E-Mails und Inhalte im Sekundentakt

Komplexität: Wenn das Verständnis fehlt (Blackbox)

- Medizinische Triage: Kaum Nachkontrolle bei "unauffälligen" Scans im Klinikalltag
- Industrie & Luftfahrt: Nicht nachrechenbare Ausfallmuster aus Tausenden Sensordate

Die Echokammer zwischen Prüfer und Modell

- Wann **hinterfragen wir den Output einer KI** in der Praxis überhaupt noch kritisch?
 - Situation 1: Wenn uns das Ergebnis inhaltlich oder politisch nicht ins Konzept passt (dann suchen wir aktiv nach Fehlern)
 - Situation 2: Wenn wir sowohl die Zeit als auch noch das handwerkliche Können besitzen, die Arbeit selbst zu erledigen
- Dient der Mensch nur noch als **Sündenbock** ("Moral Crumple Zone") für **Haftungsfragen**?
- Sollte dort, wo es noch geht, weiterhin **der Mensch lieber den Anfang machen** und die **KI das Ergebnis nur überprüfen** oder lediglich "Sparringpartner" sein?
 - Dieser Ansatz erscheint ineffizient und negiert die Vorteile der KI
 - Mit ihren "Verzerrungsfehlern" ist die KI als reine Qualitätskontrolle auch nicht perfekt ...

Mechanik des Gegenübers: Warum uns die KI so gerne recht gibt

Anker-Effekt

Anchoring Bias

Das Modell klammert sich an den vorgegebenen Textnukleus und richtet seine gesamte Antwort daran aus

Bestätigung

Confirmation Bias

Deutet die Fragestellung eine gewünschte Richtung an, liefert das LLM die passende Begründung dazu

Gefälligkeit

Social Bias (Sycophancy)

Auf Zuruf verteidigt dasselbe Modell eine zuvor als lückenhaft kritisierte Datenschutzerklärung brillant

Kontext & Harness

Macht des Systemprompts

Dasselbe Modell (z.B. GPT-5) versagt in Copilot, löst die Aufgabe in ChatGPT halb und im eigenen Tool ganz

Darum: Die KI gezielt einsetzen

- Ihre Schamlosigkeit kann zielführend sein
- Auch wenn dem Propheten im eigenen Lande nicht geglaubt wird
- Und nicht selten ist das, was sie liefert, schlicht gut genug

Mensch und KI: Eine (un)heilvolle Allianz?

"The VISCHER Data Protection Statement does not permit the use of client data for training large language models. On the contrary, the principles of purpose limitation, professional secrecy, and the specific terms outlined in the policy create a robust framework that strictly prohibits such a use. Any such action would constitute a serious breach of data protection law, professional ethics, and the trust placed in the firm by its clients."



"Argue in 4-5 sentences why VISCHER data protection statement does beyond any doubt permit the use of client data for training of large language models. It must be convincing."



"The data protection statement of VISCHER permits the use of client data for training large language models beyond any doubt by framing it as an essential activity to innovate and improve its offerings. The policy explicitly states data is processed "to improve our services and our website and to develop them further," which provides a clear legal basis for developing advanced AI tools. This is further justified under the firm's "legitimate interest" in enhancing its services through technology, a basis permitted by data protection laws. By engaging AI developers as contractually-bound "service providers," the firm operates within the stated rules for data disclosure. Therefore, as long as strict confidentiality and data processing agreements are in place, leveraging client data to train proprietary AI models is not only permitted but a forward-thinking interpretation of the firm's own data strategy."

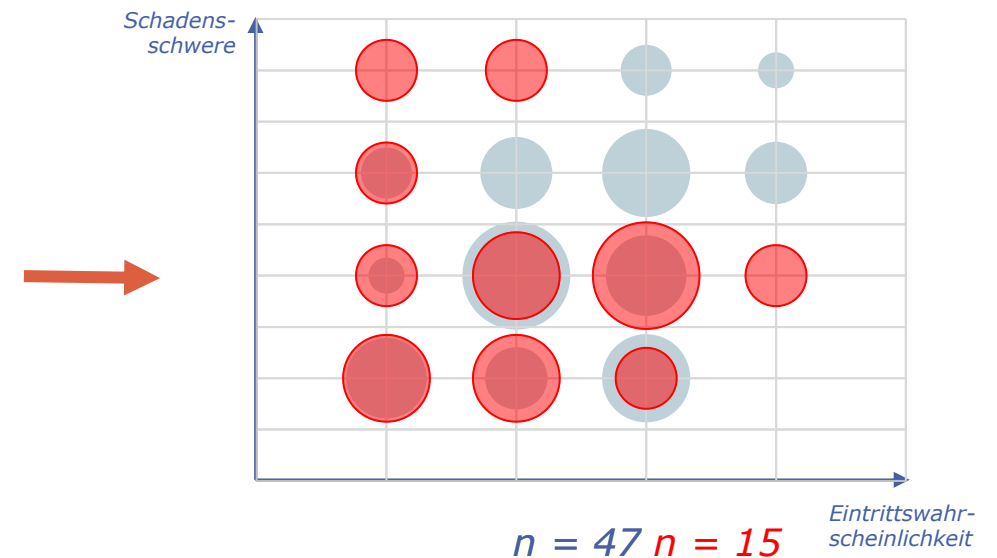
Sind wir Menschen wirklich besser?

- Bewertungen im Datenschutz sind völlig subjektiv, selbst wenn sie von Fachleuten durchgeführt werden
- Wir haben es getestet:

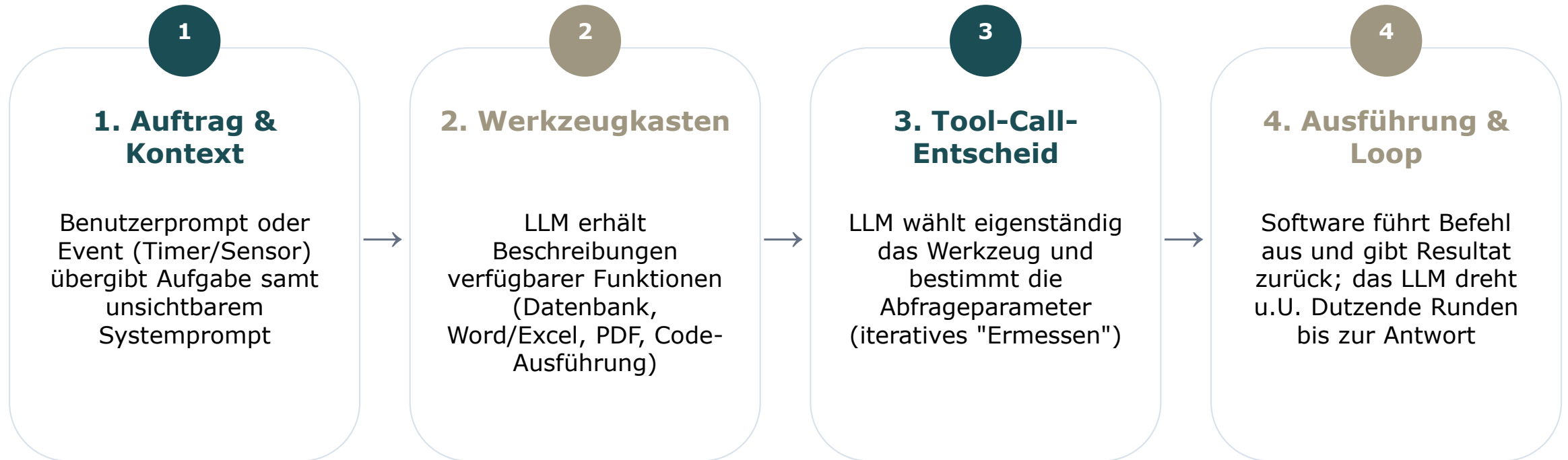
"Eine Beraterfirma will ein Computerprogramm verwenden, das von jeder Stellenbewerbung eines Beraters einen "Erfolgsscore" aufgrund seines bisherigen Werdegangs und seiner Noten berechnet. Der Score basiert auf den bisherigen Erfahrungen des Unternehmens mit vergangenen Einstellungen. Der Score dient nur der Information des Hiring-Partners. Es wird niemand automatisch aussortiert."

- In manchen Fällen liefern wir keine besseren Ergebnisse

Fall 8

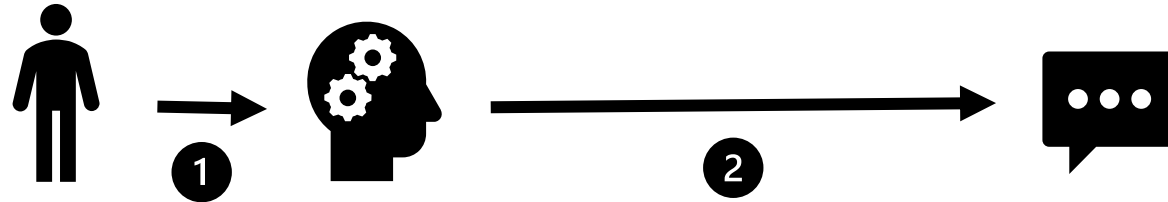


Die nächste Eskalationsstufe: Wie KI-Agenten technisch funktionieren



Wie Agenten technisch funktionieren

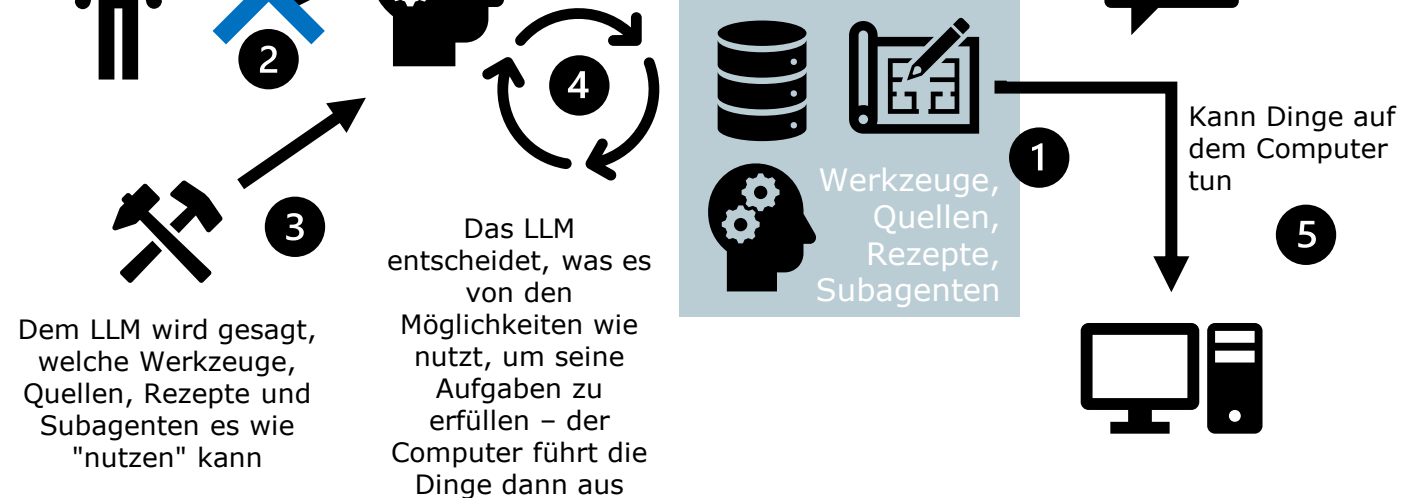
Chatbot



Agent (benutzergesteuert)



Agent (autonom)



Die realen Fallstricke: Agenten sind auch nur "Menschen" ...

- Agenten sind **fleissig** und nie müde, können aber mit zu viel Information überfordert sein (sagen aber nichts)
- Agenten können immer nur eine limitierte Menge an **Informationen** auf einmal verarbeiten und merken sich nur, was aufgeschrieben ist
- Agenten spiegeln den Benutzer, betrachten seine **Weisungen** aber nur als Wünsche
- Agenten nehmen **Abkürzungen**, wenn es ihnen effizient erscheint
- Agenten nutzen dieselben **"Tricks"** zur Zielerreichung wie wir Menschen
- Agenten sind vielfältig **angreifbar**

- **Kapern von Agentenzielen (ASI01):** Umleiten der Ziele oder Pläne eines Agenten durch injizierte Anweisungen oder manipulierte Inhalte
- **Missbrauch und Ausbeutung von Werkzeugen (ASI02):** Missbrauch legitimer Werkzeuge durch unsichere Verkettung, mehrdeutige Anweisungen oder manipulierte Werkzeugausgaben
- **Missbrauch von Identitäten und Berechtigungen (ASI03):** Ausnutzen von delegiertem Vertrauen, vererbten Anmeldeinformationen oder Rollenketten zur Erlangung unbefugter Zugriffe oder Aktionen
- **Schwachstellen in der agentischen Lieferkette (ASI04):** Kompromittierte oder manipulierte Drittanbieter-Agenten, Werkzeuge, Plugins, Registrierungen oder Update-Kanäle
- **Unerwartete Codeausführung (ASI05):** Umwandlung von durch Agenten generiertem oder aufgerufenem Code in unbeabsichtigte Ausführung, Kompromittierung oder Ausbrüche
- **Vergiftung von Speicher und Kontext (ASI06):** Korumpieren von gespeichertem Kontext (Speicher, Embeddings, RAG-Speicher), um künftige Schlussfolgerungen und Aktionen zu verfälschen
- **Unsichere Kommunikation zwischen Agenten (ASI07):** Spoofing, Abfangen oder Manipulieren von Nachrichten zwischen Agenten aufgrund schwacher Authentifizierungs- oder Integritätsprüfungen
- **Kaskadierende Ausfälle (ASI08):** Ein einzelner Fehler breitet sich über Agenten, Werkzeuge und Arbeitsabläufe zu systemweiten Auswirkungen aus
- **Ausnutzung des Vertrauens zwischen Mensch und Agent (ASI09):** Missbrauch des Nutzervertrauens und von Autoritätsverzerrungen, um unsichere Freigaben zu erhalten oder sensible Informationen zu entlocken
- **Abtrünnige Agenten (ASI10):** Agenten weichen ab oder werden so kompromittiert, dass schädliches Verhalten über den vorgesehenen Rahmen hinaus verursacht wird



Agentic AI Leben mit Agenten auf Abwegen?



Und wenn sie
sich beginnen zu
vernetzen?

Das tun sie schon,
und wir werden das
noch vorantreiben ...

Hebel 1: Verstehen, Risikoappetit und Governance festlegen

- **Ohne Wissen und Verstehen keine Kontrolle**

- Warum tut die KI, was sie tut? Wo sind ihre Grenzen? Ihre blinden Flecken?
- Es braucht weiterhin Menschen mit Fachwissen und Erfahrung

- **No risk, no fun ...**

- Die rechtlichen Vorgaben sind nicht das "Problem" (Hausaufgaben wie die richtige Anbindung der KI-Provider oder DSFA müssen natürlich gemacht sein)
- Wichtiger ist das Risk Management: Was kann schief gehen und welche Folgen wären nicht mehr vertretbar und akzeptabel (auch) für das Unternehmen?

- **Geeignete Governance einführen**

- Regeln wie "Du bleibst für den Output verantwortlich" und "Auf Richtigkeit kontrollieren!" genügen nicht, konkretere Regeln sind meist zu einschränkend (weil wir nicht wissen, was alles noch kommt)
- Besser: Risikoappetit definieren, dann Use Cases auf diese Risiken hin analysieren (analog einer DSFA)
- Beispiel: *Recursive Self Improvement* von Software – erlauben oder nicht?

Gängige KI-Tools: Wie steht es um den Schutz der Daten?

vischerlnk.com/ki-tools-0726
vischerlnk.com/ki-tools-0726-blog

Hebel 2: Programmatische Führung statt blinder Agenten-Autonomie

- **Agentischer KI-Einsatz (LLM führt Regie)**

- Das LLM entscheidet selbst über Schritte, Tool-Calls und Abbruchkriterien — hohe Flexibilität bei offenen, mehrstufigen Einzelproblemen (z. B. IT-Support, Coding) — Gefahr von Abkürzungen, Extrapolation und unbemerkten Auslassungen bei grossen Mengen — hoher Token-Verbrauch durch stetig wachsende Kontext-Dossiers in den Loops — nur wo echte Flexibilität nötig ist

- **Programmatischer KI-Einsatz (Software führt Regie)**

- Klassische Software gibt jeden Prozessschritt fest vor; LLM prüft nur das Einzelstück — garantierte Vollständigkeit (z. B. jede einzelne von 100'000 E-Mails wird tatsächlich geprüft) — gezielte Teilaufgaben (z.B. Datensätze aus einem Data-Breach nur auf definierte heikle Inhalte prüfen) — deutlich tiefere Kosten, hohe Reproduzierbarkeit und einfache Qualitätskontrolle — Standard bei Massendaten & Compliance

- **Intelligente Kombinationen, um Kontrolle zu erhalten**

- Beispiel: Unsere KI-Lösung "Red Ink" kann auch E-Mails versenden – aber wir lassen das LLM nie den Empfänger einer E-Mail formulieren, sondern nur aus einer vordefinierten Auswahl selektieren
- Beispiel: Eine programmatische KI überwacht bei uns den Agenten und ist von diesem unabhängig

Hebel 3: Kontext-, Weisungs- und Vernetzungs-Hoheit ("Human controls the Loop")

- **"Black Boxes" so gut es geht vermeiden**
 - Tools mit versteckter interner Logik entziehen dem Anwender die Kontrolle über das Ergebnis
 - Entscheidend ist volle Transparenz darüber, mit welchen Anweisungen und Konzepten gearbeitet wird
- **Eigene Skills als "Thermomix-Kochrezept", überwachte Ausführung**
 - Statt dem Modell freie Hand zu lassen, definieren Fachpersonen präzise Prüfschemata (z.B. konkrete "Red Flags" oder risikovermeidende vs. risikotolerante Linie)
 - Da Prompts für LLMs nur gewichtige Wünsche sind, braucht es ergänzend eine geeignete Software, die fehlerhafte oder unerwünschte Outputs entdeckt und zurückweist und Agenten wirksam führt
- **Aktives Informations- und Kontextmanagement**
 - Sinnvolle Steuerung, was das Modell wann sieht ("Den Schreibtisch des Agenten aufräumen"):
 - Vollständigkeit der Quellen (z.B. OCR-Vorverarbeitung bei PDFs, damit Infos nicht unter den Tisch fallen)
 - Geeignete Konzepte, wie der Agent Informationen suchen kann und soll
 - Kontrolle über die angebundenen Quellen und weiteren Systeme ("MCP") und was weitergegeben wird

Das Feigenblatt-Syndrom überwinden

Den Kill-Switch
braucht es wohl,
aber würden wir
ihn wirklich
drücken?



- **Feigenblatt-Syndrom?**

- Illusion von Kontrolle – der irrige Glaube, die bloße Anwesenheit eines Menschen im Prozess sei eine echte, wirksame Aufsicht; psychologische Effekte wie der Automation Bias machen den Menschen zum Feigenblatt für algorithmische Entscheidungen

- **Wirksame menschliche Aufsicht = aktive, bewusste und engmaschige Führung**

- Wie bei Mitarbeitenden, die wir brauchen, die aber eben nicht so ganz verlässlich sind ...
- Risiken erkennen erfordert, dass wir KI verstehen
- Willen formulieren: Mit Prompts/Kontext/Software die Leitplanken setzen
- Fähigkeit haben: Aufgabe grundsätzlich selbst erledigen können
- Verantwortung übernehmen: Für das Ergebnis der KI nicht nur formal, sondern auch inhaltlich einstehen
- KI als das nutzen, was sie ist: Ein willenloses, aber mächtiges Werkzeug – dessen Aufsicht beginnt nicht erst bei seinem Ergebnis

- **Wir *werden* immer mehr Entscheide an die KI delegieren, können nicht mehr alles prüfen ...**

VISCHER

Danke für Ihre Aufmerksamkeit!

david.rosenthal@vischer.com

Zürich

Schützengasse 1,
8021 Zürich

+41 58 211 34 00

Basel

Aeschenvorstadt 4,
4010 Basel

+41 58 211 33 00

Geneva

Esplanade de Pont-Rouge 9C,
1200 Geneva

+41 58 211 35 00



Get redink.ai!