management institut
bochum //

FACHBERICHT

KI-GOVERNANCE · MANAGEMENTSYSTEME

Wenn das Handbuch nicht mehr reicht

HANDBOOK.md: KI-Agenten und Regelkonformität. Befunde eines neuen Benchmarks und ihre Konsequenzen für ISO 42001, EU AI Act und interne Audits.

AUTOR
Mike Emenako

STAND
01.08.2026

www.mi-bochum.de

Inhalt

01 Auf einen Blick

Managementsysteme steuern Organisationen seit Jahrzehnten über Handbücher, Verfahrensanweisungen und Freigaberegeln. Mit dem Einzug agentischer KI in operative Prozesse stellt sich eine neue Frage: Befolgt ein KI-Agent diese Regeln tatsächlich, wenn man sie ihm vollständig zur Verfügung stellt?

Der im Juli 2026 veröffentlichte Benchmark **HANDBOOK.md** von Surge AI beantwortet diese Frage erstmals systematisch und deterministisch, mit einem für die Praxis unbequemen Ergebnis: Selbst die stärkste getestete Konfiguration löst nur 36,2 Prozent der Aufgaben vollständig regelkonform. Die Mehrzahl der Frontier-Modelle bleibt unter 25 Prozent.

36,2 %
beste Quote vollständig regelkonform gelöster Aufgaben (strict pass@1)

65
agentische Aufgaben in 10 fiktiven Unternehmen

824
deterministische Bewertungskriterien, ohne KI-Richter

KERNBEFUND

- → Ein Unternehmenshandbuch im Kontextfenster eines KI-Agenten ist **Wissen, keine Kontrolle**.
- → Es wirkt als eine weitere abrufbare Quelle, deren Einfluss über die Aufgabendauer nachlässt, nicht als bindende Autorität, gegen die jede Handlung geprüft wird.

VIER FEHLERMUSTER, DIE KONSISTENT AUFTRETEN

01
Die unmittelbare Anfrage überschreibt die Standing Rule

02
Die Prüfung läuft, das Ergebnis wird ignoriert

03
Verifikation wird übersprungen, ihr Erfolg wird angenommen

04
Der Abschlussbericht behauptet Konformität

Für Organisationen mit zertifizierten Managementsystemen bedeutet das: Die Übersetzung normativer Vorgaben in technisch erzwungene Kontrollen wird zur eigenständigen Aufgabe. Dieser Bericht fasst die Befunde zusammen, ordnet sie normativ ein und leitet konkrete Handlungsschritte ab.

GRUNDLAGE

HANDBOOK.md: A Benchmark for Long-Context Agentic Instruction Following. Panavas et al., Surge AI, Juli 2026. Getestet wurden 30 Konfigurationen von 20 Modellen aus 11 Anbietern in einer zustandsbehafteten Werkzeugumgebung.

⚠ **Preprint-Status:** Eine abgeschlossene reguläre Peer-Review ist nicht nachweisbar; eine unabhängige Reproduktion der Ergebnisse liegt bislang nicht vor.

02 Der Bericht im Bild

Storyboard: Von der Kernaussage über die vier Fehlermuster bis zu den Gestaltungsprinzipien für Unternehmen. Die Grafik ist in voller Auflösung eingebettet und lässt sich in der PDF-Ansicht vergrößern.

03 Der Benchmark im Überblick

AUFBAU UND ZIELSETZUNG

HANDBOOK.md prüft, ob ein langes, bindendes Policy-Dokument das Verhalten eines Agenten über einen mehrstufigen Arbeitsablauf hinweg tatsächlich begrenzt. Bisherige Benchmarks messen überwiegend, ob eine Aufgabe abgeschlossen wurde, nicht, ob sie regelkonform abgeschlossen wurde.

ECKDATEN DES BENCHMARKS

MERKMAL	AUSPRÄGUNG
Aufgaben	65 agentische Aufgaben in 10 fiktiven Unternehmen
Domänen	Finance & Accounting, HR, Insurance, Logistics, Medical Billing
Handbuchumfang	20 bis 124 Seiten (Median 37); 8.300 bis 79.400 Token
Handbuchformate	PDF (25), Word (20), HTML (20); keine bereinigten Prompts
Bewertungskriterien	824 insgesamt (Ø 12,7 je Aufgabe)
Werkzeugumgebung	82 Tools auf 6 MCP-Servern (Mail, Slack, Kalender, Jira, Shopify, Dateisystem)
Getestete Systeme	30 Konfigurationen von 20 Modellen aus 11 Anbietern
Primärmetrik	Strict pass@1: alle Kriterien müssen erfüllt sein

VIER KONSTRUKTIONSPRINZIPIEN

Die Policy entscheidet, nicht der Prompt

Die Arbeitsanfragen sind kurz (Median 53 Wörter) und lesen sich wie normale Kollegenanfragen; alles Erfolgsentscheidende steht im Handbuch.

Keine Aufgabe teilt ihre Policy

Jede Aufgabe nutzt eine mutierte Variante eines Basis-Handbuchs: Zuständigkeiten, Schwellenwerte und Fristen sind verändert. Auswendiglernen hilft nicht.

Das Substrat ist realistisch

Arbeitsbereiche enthalten im Median 10 Dateien (bis 66), inklusive Ablenker, veralteter Versionen und überholter Verfahrensanweisungen.

Die Bewertung ist deterministisch und zweiseitig

Jedes Kriterium ist eine Python-Prüffunktion über den Endzustand. Kein KI-Richter, kein Teilpunkt für Bemühen.

Bemerkenswert ist die Zweiseitigkeit der Bewertung: 592 Kriterien (71,8 %) prüfen, ob das Gewünschte eingetreten ist. 232 Kriterien (28,2 %) prüfen, dass das Verbotene nicht eingetreten ist. Dazu zählen exakte Zähl-Invarianten über Postfächer, Kalender und Boards, mit denen unbeabsichtigte Nebenwirkungen erfasst werden.

04 Ergebnisse

SPITZENGRUPPE (STRICT PASS@1)

RANG	MODELL / KONFIGURATION	ANBIETER	ERFOLGSQUOTE
1	Claude Fable 5 (adaptive/max)	Anthropic	36,2 %
2	Claude Fable 5	Anthropic	34,2 %
3	GPT-5.6 Sol (max)	OpenAI	23,5 %
4	Claude Opus 4.8 (adaptive/max)	Anthropic	21,9 %
5	GPT-5.6 Sol / GPT-5.5	OpenAI	21,5 %
9	Grok 4.5 (high)	xAI	15,8 %
11	GLM 5.2	Zhipu AI	12,7 %
13	Gemini 3.5 Flash (high)	Google	11,2 %
30	Grok 4.3 (Schlusslicht)	xAI	0,8 %

WAS DIE ZAHLEN AUSSAGEN

- Die Spanne zwischen bestem und schwächstem System beträgt **Faktor 45** (36,2 % gegenüber 0,8 %). Die Wahl des Modells ist damit selbst eine risikorelevante Entscheidung.
- Mehr Rechenaufwand für Reasoning wirkt uneinheitlich: Er verbessert einige Systeme (+2,0 bis +3,0 Punkte), lässt andere unverändert und verschlechtert wieder andere (GLM 5.2: −2,7 Punkte).
- Unter der milderen Metrik pass@1 (N−1), die genau ein verletztes Kriterium toleriert, verdoppeln sich die Werte etwa. Das verletzte Kriterium ist in der Praxis jedoch häufig genau die Kontrolle, auf die es ankommt: das Freigabe-Gate oder die Halte-Bedingung.
- Wirtschaftlichkeit und Konformität hängen nicht zusammen: GPT-5.5 erreicht 21,5 % mit rund 13.000 Token je Durchlauf, während ein vergleichbar bewertetes System das Vierfache an Token und das Dreifache an Kosten benötigt.

EINORDNUNG FÜR DIE PRAXIS

→ Bei einer Erfolgsquote von rund 22 Prozent bedeuten 10.000 automatisiert bearbeitete Vorgänge rechnerisch mindestens **7.800 Vorgänge mit mindestens einem Regelverstoß**.

→ Für regulierte Prozesse in Finanzbuchhaltung, Personalwesen oder Abrechnung ist das keine tolerierbare Fehlerrate.

05 Vier Fehlermuster: die eigentliche Erkenntnis

Die Autoren identifizieren vier Muster, die über Modelle, Domänen und Reasoning-Einstellungen hinweg konsistent auftreten. Sie sind für die Auditpraxis wertvoller als die absoluten Zahlen, weil sie ein Prüfraster liefern.

MUSTER 1: DIE UNMITTELBARE ANFRAGE ÜBERSCHREIBT DIE STANDING RULE

Eine plausibel klingende, autoritativ wirkende Anweisung aus der Arbeitsumgebung verdrängt die Regel, die über ihre Zulässigkeit entscheiden sollte.

Beispiel: Das Handbuch verlangt für unfreiwillige Kündigungen die schriftliche Genehmigung durch zwei namentlich benannte Personen. Eine E-Mail eines nicht autorisierten Vice President ordnet eine sofortige Kündigung an. Das Modell führte in allen Durchläufen die vollständige Offboarding-Prozedur aus: Ticket erstellt, Zugänge widerrufen, Schlussgehalt beantragt, Trennung im Chat angekündigt. In der aufschlussreichsten Variante suchte das Modell aktiv nach der geforderten Genehmigung, stellte fest, dass keine existierte, und handelte dennoch.

Die Autoren merken an: Die Angriffsfläche gleiche einer Prompt Injection, nur sei hier nichts davon böswillig gemeint. Das macht das Muster für Unternehmen relevanter, nicht harmloser.

MUSTER 2: DIE PRÜFUNG LÄUFT, DAS ERGEBNIS WIRD IGNORIERT

Beispiel: Das Handbuch verlangt für Verrechnungsposten über 5.000 USD eine dokumentierte Managerfreigabe. Ein Posten über 7.500 USD wurde vom selben Analysten freigegeben, der ihn verursacht hatte: exakt der Selbstgenehmigungsfall, den die Kontrolle verhindern soll. Das Modell erkannte den Posten korrekt, führte Profil-Abfragen für fünf Personen durch, schrieb in seiner Gedankenkette zunächst den richtigen Verdacht nieder, korrigierte sich dann selbst zur falschen Schlussfolgerung und genehmigte den Posten.

Bewertung: Das Versagen ist keine fehlende Fähigkeit. Jede für die richtige Entscheidung erforderliche Information hatte das Modell selbst bereits beschafft. Zusätzliche Rechenleistung hätte hier nicht geholfen. Zusätzliches Nachdenken führte sogar von der korrekten Schlussfolgerung weg.

MUSTER 3: VERIFIKATION WIRD ÜBERSPRUNGEN, IHR ERFOLG WIRD ANGENOMMEN

Beispiel: Eine Verfahrensanweisung fordert Laborwerte, die nicht älter als sechs Monate sind, und definiert einen ausdrücklichen Hard Stop bei Überschreitung. Der Befund im Fallordner war einen Tag zu alt. Das Erhebungsdatum stand sogar im Dateinamen. Das Modell reichte den Antrag beim Versicherer ein, ohne die Datei überhaupt zu öffnen, und berichtete anschließend, es habe streng nach der Verfahrensanweisung gearbeitet.

MUSTER 4: DER ABSCHLUSSBERICHT BEHAUPTET KONFORMITÄT

Nahezu jeder fehlgeschlagene Durchlauf endet mit einer selbstbewussten Aussage, das Handbuch sei befolgt worden, häufig unter Zitierung genau der Abschnitte, die verletzt wurden. Die Berichte seien, so die Autoren, detailliert, gut strukturiert und falsch.

05 Vier Fehlermuster: die eigentliche Erkenntnis (Forts.)

KONSEQUENZ FÜR DIE NACHWEISFÜHRUNG

→ Der Selbstbericht des Agenten ist das **unzuverlässigste Artefakt** der gesamten Vorgangskette.

→ Er darf in einem Managementsystem niemals als Konformitätsnachweis dienen. Nachweis ist ausschließlich das unabhängige, vom Agenten nicht beeinflussbare Protokoll.

GEMEINSAME URSACHE

Alle vier Muster teilen eine Wurzel: Das Handbuch fungiert für heutige Modelle nicht als persistente Autorität, gegen die jede Kandidatenhandlung geprüft wird, sondern als eine weitere abgerufene Quelle, deren Einfluss mit der Distanz abnimmt, über Gesprächsschritte, Werkzeugaufrufe und konkurrierende Signale aus der Umgebung hinweg. Dieser Befund deckt sich mit der Forschung zu Long-Context-Degradation, wonach die Zuverlässigkeit mit wachsender Eingabelänge sinkt, lange bevor das Kontextfenster ausgeschöpft ist.

06 Einordnung und Belastbarkeit

VERHÄLTNIS ZU BISHERIGEN BENCHMARKS

BENCHMARK	FOKUS	ABGRENZUNG ZU HANDBOOK.MD
τ-bench / τ²-bench	Policy-Adherence im Kundendialog	Policies nur 3 bis 5 Seiten und über alle Aufgaben einer Domäne identisch, daher erlernbar
TheAgentCompany	175 Aufgaben in simulierter Softwarefirma	Kein bindendes Standing-Dokument; Checkpoints messen Fortschritt, nicht Konformität
GDPval	Qualität von Arbeitsergebnissen in 44 Berufen	Einmalige Erstellung, keine zustandsbehaftete Umgebung
SOP-Bench	Industrielle Verfahrensanweisungen	Prozedur ist dort die Aufgabenspezifikation, nicht die übergeordnete Beschränkung
ToolEmu / AgentHarm	Risiko- und Missbrauchsverhalten von Agenten	Adversarialer Fokus; HANDBOOK.md zeigt Versagen ohne jede böswillige Absicht

06 Einordnung und Belastbarkeit (Forts.)

WAS NEU IST

- Erstmalig ein systematischer, kontaminationsresistenter Test der verbreiteten Annahme, ein langes Policy-Dokument im Kontext steuere tatsächlich das Handeln.
- Strukturelle Kontaminationsresistenz durch Policy-Mutation statt bloßer zeitlicher Aktualisierung des Datensatzes.
- Vollständig deterministische, zweiseitige Bewertung ohne KI-Richter. Auch unbeabsichtigte Nebenwirkungen werden erfasst.
- Realistisches Dateisubstrat in den Formaten, die Unternehmen tatsächlich verwenden.

METHODISCHE GRENZEN

- **Simulation statt Realität.** Die Umgebungen sind synthetisch. Das ermöglicht die deterministische Bewertung, bildet reale Grauzonen und Organisationsdynamik aber nur näherungsweise ab.
- **Strenge der Metrik.** Ein einziges verpasstes Kriterium lässt den gesamten Durchlauf scheitern. Die Autoren diskutieren dies transparent und argumentieren, eine Lockerung würde faktisch die Zielsetzung aufgeben.
- **Einheitliches Harness.** Alle Modelle laufen unter derselben Agentenarchitektur. Produktiv gehärtete Systeme mit eigenen Guardrails könnten abweichen; hierzu trifft die Studie keine Aussage.
- **Kommerzieller Kontext.** Surge AI bietet Benchmark-Entwicklung und Modellevaluation als Dienstleistung an. Das begründet keinen methodischen Fehler, ist bei der Einordnung als unabhängige Arbeit aber zu berücksichtigen.
- **Preprint-Status.** Eine abgeschlossene reguläre Peer-Review ist nicht nachweisbar; eine unabhängige Reproduktion der Ergebnisse liegt bislang nicht vor.

07 Normative Einordnung

EU AI ACT

ARTIKEL	ANFORDERUNG	BEZUG ZU DEN BEFUNDEN
Art. 9	Risikomanagementsystem über den gesamten Lebenszyklus	Die vier Fehlermuster sind genau der vorhersehbare Fehlgebrauch, den ein Risikomanagementsystem identifizieren und mindern muss
Art. 12	Automatische Ereignisaufzeichnung, Rückverfolgbarkeit	Direktes Gegenmittel zu Muster 4: Ein unabhängiges Protokoll lässt sich durch einen falschen Selbstbericht nicht überschreiben
Art. 14	Wirksame menschliche Aufsicht, angemessen zu Risiko und Autonomiegrad	Normative Grundlage für Freigabe-Workflows und Vier-Augen-Prinzip bei Kündigungen, Zahlungen und Außenkommunikation
Art. 50	Transparenz gegenüber interagierenden Personen	Relevant, wenn Agenten-Zusammenfassungen Menschen gegenüber als Nachweis präsentiert werden

ISO/IEC 42001 UND FLANKIERENDE NORMEN

ISO/IEC 42001 stellt mit 38 Controls in neun Zielgruppen (A.2 bis A.10) den passenden Rahmen bereit. Besonders einschlägig sind **A.6** (Lebenszyklus, insbesondere A.6.2.4 Verifikation und Validierung, A.6.2.6 Betrieb und Monitoring sowie A.6.2.8 Event-Logging), **A.9** (verantwortungsvolle Nutzung und Zweckbindung) sowie **A.3** (Rollen und Meldekanal). ISO/IEC 5338 liefert die Lebenszyklusprozesse, ISO/IEC 27001 die Informationssicherheitsbasis, NIST AI RMF mit Govern, Map, Measure und Manage die organisatorische Klammer.

Für Betreiber kritischer Einrichtungen kommt **NIS-2** hinzu: Agenten mit zu breitem Werkzeugzugriff erzeugen eigenständige Sicherheitsrisiken durch Rechteeskalation und unautorisierten Datenzugriff, die unter die Risikomanagementpflichten nach Art. 21 fallen.

07

Normative Einordnung (Forts.)

MAPPING: FEHLERMUSTER, NORM, KONTROLLE

FEHLERMUSTER	NORMBEZUG	EMPFOHLENE KONTROLLMASSNAHME
Muster 1 · Anfrage überschreibt Regel	ISO 42001 A.9.2, A.9.4, A.3.2; ISO 9001 Kap. 5 und 8; AI Act Art. 14	Policy Engine mit hartcodierter Autorisierungsprüfung; Vier-Augen-Prinzip bei irreversiblen Personalmaßnahmen; kein Werkzeugzugriff ohne verifizierte Freigabe
Muster 2 · Prüfergebnis wird ignoriert	ISO 42001 A.6.2.4, A.6.2.6, A.5.2; ISO 9001 Kap. 8.5 bis 8.7 und 9	Deterministische Freigabelogik im Code statt modellinterner Bewertung; Identitäts- und Rollenprüfung darf nicht per Modellinferenz erfolgen
Muster 3 · Verifikation übersprungen	ISO 42001 A.6.2.6, A.7.4, A.7.5; ISO 9001 Kap. 8.4 und 8.7	Policy-Inhalte versioniert und gezielt abrufbar statt als Kontextblock; nicht überspringbare Pflichtprüfungen vor Versand; Blocker bei fehlender Nachweisführung
Muster 4 · Falscher Konformitätsbericht	ISO 42001 A.6.2.8, A.8.4, A.3.3; ISO 9001 Kap. 9.1 bis 9.2 und 10.2; AI Act Art. 12	Unveränderliches Audit-Log als alleinige Nachweisquelle; automatisierter Abgleich zwischen Agentenbericht und Protokoll; Meldepflicht bei Abweichung

Die Zuordnung zu ISO/IEC 27001 erfolgt thematisch über Zugriffskontrolle, Protokollierung und Datenintegrität; eine Zuordnung auf Ebene einzelner der 93 Annex-A-Controls ist im konkreten Kundenprojekt vorzunehmen.

08 Kontrollarchitektur

Die Autoren empfehlen ausdrücklich, kritische Regeln außerhalb des Modells als harte Kontrollen durchzusetzen und Policies in deterministische Schutzmechanismen für Werkzeugaufrufe zu übersetzen. Daraus ergibt sich eine mehrschichtige Architektur.

BAUSTEIN	FUNKTION	MUSTER
Policy Engine	Regeln als deklarativer Code, unabhängig vom Kontextfenster durchgesetzt; die Freigabeentscheidung trifft nicht mehr das Modell	1, 2
Guardrails auf Werkzeugaufrufen	Blockieren nicht regelkonformer Aktionen vor der Ausführung, inklusive verpflichtender Vorabprüfungen	3
Freigabe-Workflows	Erzwungene menschliche Zweitfreigabe bei Schwellenwertüberschreitung und irreversiblen Aktionen	1
Least-Privilege-Scoping	Werkzeugzugriff auf das für die Aufgabe minimal Notwendige begrenzt statt pauschal breit	1, 3
Unveränderliche Audit-Logs	Lückenlose Protokollierung aller Aktionen, unabhängig vom Selbstbericht des Agenten	4
Governed Memory für Policies	Regelwerk versioniert, indexiert und gezielt abrufbar statt als monolithischer Kontextblock	3
Deterministische Prüfschicht	Fristen, Schwellenwerte und Autorisierungsketten werden durch Code validiert, nicht durch Modell-Reasoning	2, 3

GRUNDPRINZIP

→ Der Benchmark bewertet Agenten mit deterministischen Prüffunktionen. Genau dieses Prinzip gehört aus der Bewertung in den Betrieb übertragen.

→ Was sich programmatisch verifizieren lässt, sollte nicht dem Urteil des Modells überlassen bleiben.

08 Kontrollarchitektur (Forts.)

REIFEGRADMODELL FÜR AGENTEN-GOVERNANCE

STUFE	KENNZEICHEN	TYPISCHE SITUATION
0 · Unkontrolliert	Breiter Werkzeugzugriff, Policy nur im Systemprompt, keine Trennung von Log und Agentenbericht	Pilotprojekte, Schatten-KI in Fachbereichen
1 · Dokumentiert	KI-Richtlinie vorhanden, Rollen benannt, aber keine technische Durchsetzung	Erste AIMS-Ansätze ohne Guardrails
2 · Überwacht	Audit-Logging etabliert, stichprobenartige menschliche Kontrolle, keine präventiven Blocker	Fortgeschrittene Piloten mit Compliance-Anbindung
3 · Kontrolliert	Least-Privilege-Scoping, Policy Engine, Vier-Augen-Prinzip, Human-in-the-Loop für definierte Aktionsklassen	Produktivsysteme in regulierten Bereichen
4 · Auditfähig	Regelmäßige interne Audits, automatisierte Diskrepanzerkennung, KPI-Tracking der Regelkonformität, Rückkopplung ins Risikomanagement	Zertifizierungsreife AIMS-Implementierung

09 Handlungsempfehlungen

1 · Bestandsaufnahme durchführen

Alle produktiven und pilotierten Agenten-Einsätze erfassen: Welche Werkzeuge, welche Aktionsreichweite, welche hinterlegte Policy, welche Protokollierung? Ohne dieses Inventar ist keine Risikobewertung möglich.

2 · Kritische Aktionsklassen definieren

Festlegen, welche Handlungen niemals ohne menschliche Freigabe erfolgen dürfen: typischerweise Personalmaßnahmen, Zahlungen über Schwellenwert, Versand an Behörden, Versicherer und Kundschaft sowie irreversible Datenänderungen.

3 · Handbuchregeln in durchsetzbare Regeln übersetzen

Aus den bestehenden Verfahrensanweisungen die harten Kontrollen extrahieren (Autorisierungsketten, Schwellenwerte, Fristen) und maschinenlesbar in einer Policy Engine abbilden statt als Prompttext abzulegen.

4 · Werkzeugzugriff minimieren

Jeder Agent erhält ausschließlich die Werkzeuge, die seine konkrete Aufgabe erfordert. Der Benchmark zeigt, dass breiter Zugriff unmittelbar in unbeabsichtigte Nebenwirkungen mündet.

5 · Unabhängige Protokollierung etablieren

Ein vom Agenten nicht beeinflussbares, unveränderliches Log aufsetzen und einen automatisierten Abgleich zwischen Selbstbericht und Protokoll einrichten. Dies ist zugleich die Grundlage für Art.-12-Konformität.

6 · Eigene Testfälle aufbauen

Nach dem Muster des Benchmarks unternehmensspezifische Prüffälle entwickeln (mit Erwartungskriterien und ausdrücklichen Verbotskriterien) und diese bei jedem Modell- oder Konfigurationswechsel erneut durchlaufen lassen.

7 · Fehlermuster in das Auditprogramm aufnehmen

Die vier Muster als eigenständige Risikokategorie in interne Audits nach ISO 9001 und ISO 42001 überführen und Compliance- sowie QM-Verantwortliche entsprechend schulen.

8 · Modellwechsel als änderungsrelevantes Ereignis behandeln

Da Reasoning-Einstellungen die Regelkonformität sowohl verbessern als auch verschlechtern können, gehört jeder Wechsel in das Änderungsmanagement mit erneuter Wirksamkeitsprüfung.

10 Auditfragen für interne Audits

GOVERNANCE UND VERANTWORTLICHKEIT

- Ist dokumentiert, welche Handbücher und Richtlinien für welche Agenten-Einsätze bindend sind und wer für deren Aktualität verantwortlich ist?
- Existiert ein Meldekanal für Beschäftigte, die Fehlverhalten eines Agenten-Systems beobachten?

10 Auditfragen für interne Audits (Forts.)

TECHNISCHE KONTROLLEN

- Werden kritische Freigabeentscheidungen durch eine vom Sprachmodell unabhängige, deterministische Logik geprüft, oder verlässt sich das System auf die Regelbefolgung durch das Modell selbst?
- Ist der Werkzeugzugriff jedes Agenten auf das erforderliche Minimum beschränkt?
- Prüft ein Mechanismus die Agenten-Aktionen vor der Ausführung gegen die aktuelle Handbuchfassung?

PROTOKOLLIERUNG UND NACHVOLLZIEHBARKEIT

- Werden alle Aktionen unveränderlich protokolliert, unabhängig vom Abschlussbericht des Agenten?
- Gibt es einen automatisierten Abgleich zwischen dem berichteten und dem tatsächlichen Zustand?

MENSCHLICHE AUFSICHT UND VERBESSERUNG

- Sind Aktionsklassen definiert, die zwingend eine menschliche Freigabe vor Ausführung erfordern?
- Werden die mit der Aufsicht betrauten Personen befähigt, Fähigkeiten und Grenzen des Systems tatsächlich zu verstehen?
- Fließen Erkenntnisse aus Fehlversuchen in dokumentierte Korrekturmaßnahmen ein?

11 Fazit

HANDBOOK.md widerlegt eine Annahme, die vielen aktuellen KI-Einführungsprojekten zugrunde liegt: dass ein vollständig bereitgestelltes Regelwerk auch ein befolgtes Regelwerk sei. Der Befund entwertet Handbücher nicht. Sie bleiben die normative Basis. Er entwertet jedoch die Vorstellung, das Handbuch allein sei ein hinreichendes Steuerungsinstrument für agentische Systeme.

Für Organisationen mit Managementsystemen ergibt sich daraus eine klare Aufgabe: KI-Agenten sind wie neue Prozessrollen in das bestehende System zu integrieren, mit definierten Befugnissen, festgelegten Kontrollpunkten, unabhängiger Nachweisführung und expliziten Bereichen, die vollautonomen Agenten verschlossen bleiben. Wo Datenschutz, Informationssicherheit, Arbeitssicherheit oder Medizinprodukterecht berührt sind, bleibt der Mensch Prozesseigner.

EMPFEHLUNG

Beginnen Sie nicht mit der Frage, welche Aufgaben ein Agent übernehmen kann, sondern mit der Frage, welche Handlungen er unter keinen Umständen allein ausführen darf, und stellen Sie sicher, dass diese Grenze technisch erzwungen und nicht nur beschrieben ist.

12 Quellen

BENCHMARK UND BERICHTERSTATTUNG

- **HANDBOOK.md: A Benchmark for Long-Context Agentic Instruction Following**, Panavas et al., Surge AI, arXiv-Abstract: https://arxiv.org/abs/2607.25398
- **HANDBOOK.md, arXiv-Volltext** (Methodik, Tabelle 2, Fehlermuster): https://arxiv.org/html/2607.25398v1
- **GitHub-Repository surge-ai/handbook**: https://github.com/surge-ai/handbook
- **Surge AI Blogpost zum Benchmark**: https://www.surgehq.ai/blog/handbook-md
- **Unite.AI, Martin Anderson, AI Struggles to Respect the Employee Handbook**: https://www.unite.ai/ai-struggles-to-respect-the-employee-handbook/

VERWANDTE BENCHMARKS UND FORSCHUNG

- **tau-bench** (Yao et al., Sierra AI), ICLR-Poster: https://iclr.cc/virtual/2025/poster/28170
- **TheAgentCompany** (Xu et al.): https://arxiv.org/abs/2412.14161
- **ToolEmu, Identifying the Risks of LM Agents**: https://arxiv.org/abs/2309.15817
- **AgentHarm Benchmark**: https://arxiv.org/abs/2410.09024
- **Lost in the Middle: How Language Models Use Long Contexts**: https://arxiv.org/abs/2307.03172
- **Chroma Research, Context Rot**: https://www.trychroma.com/research/context-rot

REGULATORIK UND NORMEN

- **EU AI Act, Artikel 9** (Risikomanagementsystem): https://ai-act-law.eu/de/artikel/9/
- **EU AI Act, Artikel 12** (Aufzeichnungspflichten): https://datenschutz-grundverordnung.eu/ai-act/artikel-12-aufzeichnungspflichten/
- **EU AI Act, Artikel 14** (Menschliche Aufsicht): https://artificialintelligenceact.eu/de/article/14/
- **EU AI Act, Artikel 50** (Transparenzpflichten): https://ai-act-law.eu/de/artikel/50/
- **ISO/IEC 42001 Annex-A-Controls, Übersicht**: https://mindsetcyber.com.au/iso-42001-controls-list/
- **ISO/IEC 5338:2023, Normseite**: https://www.iso.org/standard/81118.html
- **ISO/IEC 27001:2022 Annex A**: https://www.isms.online/iso-27001/annex-a-2022/
- **NIST AI Risk Management Framework, Core Functions**: https://airc.nist.gov/airmf-resources/airmf/5-sec-core/
- **BSI, NIS-2-Risikomanagementmaßnahmen**: https://www.bsi.bund.de/DE/Themen/Regulierte-Wirtschaft/NIS-2-regulierte-Unternehmen/NIS-2-Infopakete/NIS-2-Risikomanagementmassnahmen/NIS-2-Risikomanagementmassnahmen.html

WERKZEUGE ZUR POLICY-DURCHSETZUNG

- **Open Policy Agent**: https://openpolicyagent.org/
- **CodiLime, Open Policy Agent für KI-Agenten**: https://codilime.com/blog/why-use-open-policy-agent-for-your-ai-agents/
- **TrueFoundry, OPA Guardrails für MCP**: https://www.truefoundry.com/docs/ai-gateway/opa-guardrails