

Der Mensch setzt die Regeln

Vier Ebenen der Aufsicht: wo die verantwortliche Fachperson steht, wenn Agenten arbeiten und prüfen

Falk Uebernickel

Universität St.Gallen (HSG)

Menschlich gedacht, mit KI geschrieben: Ideen und Argumentation stammen vom Autor; der Text entstand gemeinsam mit KI.

Zusammenfassung. „Human-in-the-Loop“ klärt, *ob* Menschen KI beaufsichtigen, aber nicht, auf welcher Ebene. Dieses Arbeitspapier schlägt vier Ebenen der Aufsicht vor: die Arbeit, das Ergebnis, dessen Prüfung und die Regeln, die festlegen, was als richtig gilt. Über drei Stufen der KI-Reife hinweg, die in den meisten Organisationen nebeneinander bestehen, steht die verantwortliche Fachperson jeweils auf einer anderen Ebene. Gestern beaufsichtigten Fachleute Menschen und sahen die Arbeit. Heute erstellen Agenten in großer Menge, und die Fachperson gibt jedes Ergebnis einzeln frei; nach unserer Beobachtung stehen hier die meisten Organisationen, die KI in großem Umfang einsetzen, und hier sitzt der Engpass. Morgen, als Hypothese, prüfen Aufsichts-Agenten jedes Ergebnis gegen explizite Regelsets, und die Fachperson geht eine Ebene höher: Sie schreibt die Regeln und prüft die Prüfer. Wir legen dar, was dieser Schritt voraussetzt und wo er endet, wie sich Aufgabe und Erfolgsmaß der Fachperson verändern, und ziehen fünf Folgerungen für Führungskräfte. Das Arbeitspapier knüpft an Uebernickel (2026) an.

1 Vom Ob zum Wo

„Human-in-the-Loop“ beruhigt und bleibt unscharf. Wer ein Ergebnis überfliegt, bevor es das Haus verlässt, ist im Loop; ebenso die Fachperson, die entscheidet, was ein Agent überhaupt tun darf. Die Formel deckt beides ab und schweigt über den Unterschied, dabei entscheidet er darüber, ob die Aufsicht hält, wenn das Volumen wächst.

Im Arbeitspapier *Jede Aufgabe lässt sich mit KI unterstützen* (Uebernickel, 2026) haben wir argumentiert, dass die nützliche Frage zu KI in der Arbeitswelt nicht lautet, *ob* sich eine Aufgabe unterstützen lässt, sondern *wie*, und dass die Form der Unterstützung davon abhängt, wie sich das Ergebnis prüfen lässt. Folgt die Form der Unterstützung der Prüfbarkeit, dann muss mit günstigerer Prüfung auch der Ort der menschlichen Aufsicht wandern. Die Frage wird zum *Wo*: Wie weit von der Arbeit steht die verantwortliche Fachperson, und wofür ist sie verantwortlich?

2 Vier Ebenen der Aufsicht

Wir unterscheiden vier Ebenen, geordnet nach ihrem Abstand zur Arbeit selbst:

- **Ebene 1 • Arbeit.** *Wie wurde es gemacht?* Wer beaufsichtigt, sieht den Prozess: Quellen, Schritte, Ermessensentscheidungen.
- **Ebene 2 • Ergebnis.** *Was wurde geliefert?* Wer beaufsichtigt, entscheidet, ob es hinausgeht.
- **Ebene 3 • Prüfung.** *Stimmt das Ergebnis?* Wer beaufsichtigt, prüft das Ergebnis gegen seine Quelle und gegen die geltenden Kriterien. Hier sind Qualität und Richtigkeit verortet.
- **Ebene 4 • Regeln.** *Was gilt überhaupt als richtig?* Wer beaufsichtigt, setzt die Kriterien. Es empfiehlt sich, zwei Regelsets zu unterscheiden: *Muss-Regeln* sagen, was nie

GEGENPOSITION

Fachleute können womöglich nicht aufschreiben, was sie wissen. Nisbett und Wilson (1977) trugen Befunde zusammen, wonach Menschen kaum introspektiven Zugang zu den Prozessen hinter ihren Urteilen haben und stattdessen plausible Theorien berichten. Das Regelset einer Fachperson beschreibt womöglich eine Theorie der Arbeit statt der Arbeit selbst. Welche Regeln werden aufgeschrieben: die, die das Urteil leiten, oder die, die so klingen?

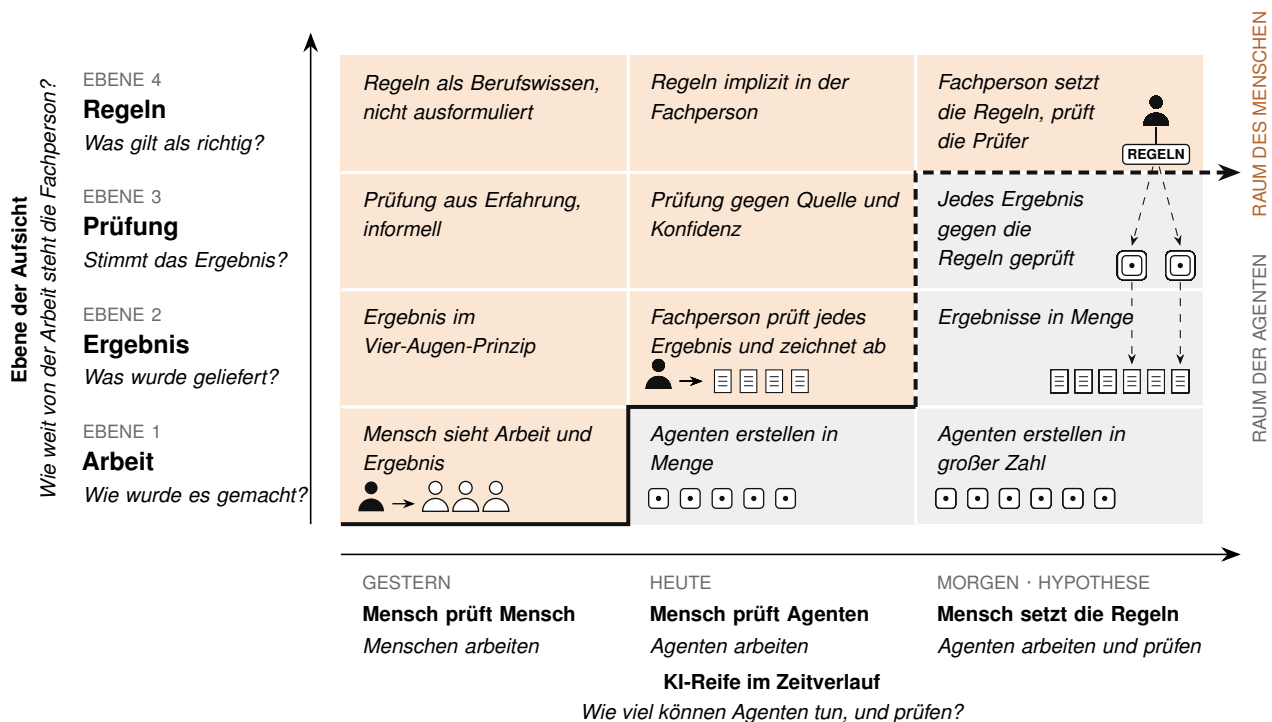


Abbildung 1. Vier Ebenen der Aufsicht über drei Stufen der KI-Reife. Dicke Linie: Position der verantwortlichen Fachperson; darunter haben Agenten übernommen. Dünne Linie zum Regelset: setzt die Regeln; gestrichelte Pfeile: Prüfungen der Aufsichts-Agenten. Gestrichelt: Hypothese; setzt vorab festgelegte Kriterien voraus (Frontloading). Gefüllte Figur: Fachperson; Kontur: ausführende Menschen; Quadrate: Agenten; doppelt gerahmt: Aufsichts-Agenten. Eigener Vorschlag; Einzelfreigabe und Frontloading nach Hörter (2026).

erlaubt ist (Recht, Compliance, vertragliche Pflichten); *Soll-Regeln* sagen, was die Führung tatsächlich will (Preiskorridore, Segmentfokus, No-Gos in der Positionierung, Eskalationsschwellen).

Jede Ebene bezieht ihren Maßstab von der Ebene darüber: Eine Prüfung ohne explizite Regeln prüft gegen das, was die prüfende Person gerade im Kopf hat. Ebene 4 sagt, was als richtig gilt, Ebene 3, ob es im Einzelfall hält (Abbildung 1).

3 Drei Stufen: wo die Fachperson steht

Die Spalten von Abbildung 1 sind Stufen, keine Zeitpunkte: Sie bestehen nebeneinander, oft im selben Unternehmen.

Gestern: Mensch prüft Mensch. Fachleute beaufsichtigten Menschen: Sie sahen Arbeit und Ergebnis, wandten das Vier-Augen-Prinzip an und urteilten aus Erfahrung. Die Regeln waren Berufswissen, weitgehend ungeschrieben; bei menschlichem Volumen genügte das.

Heute: Mensch prüft Agenten. Agenten erstellen in großer Menge, und die Fachperson prüft jedes Ergebnis gegen seine Quelle und zeichnet es ab; sie vereint damit die Ebenen 2 und 3 in einem Akt: das Muster der Einzelfreigabe nach Hörter (2026). Nach unserer Beobachtung stehen hier die meisten Organisationen, die KI in großem Umfang einsetzen, und hier sitzt der Engpass, denn die Zeit der Fachleute skaliert nicht mit der Zahl der Ergebnisse. Zwei Schwächen folgen. Die Regeln sitzen im Kopf der Fachperson; sie lassen sich weder teilen noch versionieren noch anfechten. Und wer überwiegend korrekte Ergebnisse sieht, verlässt sich auf sie und übersieht den seltenen Fehler (Parasuraman und Riley, 1997; Skitka et al., 1999); die Freigabe driftet zum Abnicken.

Morgen: Mensch setzt die Regeln (Hypothese). Aufsichts-Agenten prüfen jedes Ergebnis gegen seine Quelle und gegen zwei explizite Regelsets, Muss-Regeln und Soll-Regeln; im hier vorgeschlagenen Entwurf hat jedes Regelset seinen eigenen Aufsichts-Agenten.

GEGENPOSITION

Das Urteil von Fachleuten ist ein schwacher Maßstab, auf jeder Ebene. In rund hundert Vergleichen fanden Dawes et al. (1989) einfache aktuarische Regeln dem klinischen Expertenurteil fast durchweg ebenbürtig oder überlegen; durften Fachleute sie übersteuern, sank die Genauigkeit eher. Wenn Prüfer und Fachperson uneins sind: Wer hat eher recht?

Die Fachperson rückt auf Ebene 4: Sie schreibt die Regeln und prüft die Prüfer, indem sie Ergebnisse mit eigenem Urteil stichprobenartig prüft, die Urteile der Prüfer mit bekannten Antworten vergleicht und auf Drift achtet. Schriftliche Prinzipien leiten bereits heute die Kritik von KI an KI im Modelltraining (Bai et al., 2022). Die Spalte „Morgen“ ist das organisationale Gegenstück dessen, was die Alignment-Forschung *Scalable Oversight* (skalierbare Aufsicht) nennt (Bowman et al., 2022).

4 Was der Schritt voraussetzt, und wo er endet

Frontloading. Der Schritt setzt Frontloading voraus: Kriterien, die feststehen, bevor etwas erzeugt wird, ein Muster, das Hötter (2026) im Vertrieb von Unternehmenssoftware beobachtet hat. Wo der Beleg noch nicht existiert, wie bei einem ROI-Versprechen, gibt es keine Regel, gegen die sich prüfen ließe; solche Aufgaben bleiben in der Variante „Evidenzgeführt über die Zeit“ aus *Jede Aufgabe lässt sich mit KI unterstützen*.

Umkehrbarkeit. Wo sich eine Wirkung nicht rückgängig machen lässt, bleibt die Schranke vor der Wirkung menschlich, was immer die Prüfer sagen. Autonomie sollte nie schneller laufen als die Fähigkeit einzugreifen, wie Carlos O. Fuentes es in einem Kommentar zu *Jede Aufgabe lässt sich mit KI unterstützen* formuliert hat.

Unabhängigkeit der Prüfer. Ein Aufsichts-Agent aus derselben Modellfamilie wie der Generator teilt womöglich dessen blinde Flecken. Studien zu Sprachmodellen als Richtern belegen Positions- und Längeneffekte und, schwächer belegt, eine Neigung zur Selbstbevorzugung (Zheng et al., 2023). Prüfer sollten einfacher und unabhängiger sein als die Generatoren: regelbasiert, wo Regeln genügen, sonst auf anderem Modell oder anderen Daten.

Das Abnicken wandert mit nach oben. Jede Ebene erbt das Fehlerbild der Ebene darunter. Wer die Urteile der Prüfer stichprobenartig sichtet, kann dazu übergehen, sie durchzuwinken; Ebene 4 braucht dieselben Messgrößen wie Ebene 2 (Abschnitt 5).

Wo das Regelwissen entsteht. Wer eine gute Regel schreiben will, muss wissen, was schiefeht, und das lernt man in der Arbeit, die delegiert wird. Bainbridge (1983) nannte dies eine Ironie der Automatisierung: Die Fähigkeit, ein System zu überwachen, verkümmert, sobald das System die Arbeit erledigt. Fachleute für Ebene 4 fit zu halten heißt, einen bewussten Anteil eigener Arbeit zu bewahren.

Die Spezifikationslücke. Agenten prüfen gegen Regeln, nicht gegen Absichten; eine Regel kann erfüllt und ihr Zweck dennoch verfehlt sein. Absicht in Regeln zu übersetzen und zu bemerken, wo beide auseinanderlaufen, wird zur Kernaufgabe der Fachperson. In Tim Hourigans Unterscheidung sind das Steuern von Verhalten und das Steuern von Zweck zwei verschiedene Aufgaben; nur die erste wandert zur Maschine.

5 Die Aufgabe der Fachperson verändert sich

Die drei Stufen beschreiben eine Rollenleiter. Gestern beaufsichtigte die Fachperson die Menschen, die die Arbeit taten, und beurteilte das Ergebnis; Erfolg war das abgelieferte Ergebnis. Heute gibt sie frei, und Erfolg sollten die gefundenen Fehler sein, nicht der Durchsatz an Freigaben. Morgen setzt sie die Regeln und prüft die Prüfer, und Erfolg ist, was die Regeln fangen und was noch durchrutscht. Eine Ebene höher zu gehen bedeutet weniger Fachleute je Ergebnis, nicht unbedingt weniger Fachleute: Die Arbeit wird zum Schreiben von Regeln, zum Kalibrieren von Prüfern und zum Entscheiden von Ausnahmen.

In der Praxis werden Rollen häufiger ausgehöhlt als neu gestaltet. Die Arbeit wandert zum Freigeben, während Stellenprofil, Anreize und Laufbahn beim Erstellen bleiben. Drei Messgrößen unterscheiden Aufsicht von Abnicken, auf jeder Ebene: die Ablehnungsquote (nahe null ein Warnsignal), das Auffinden eingestreuter Fehler und wie oft ein blindes

GEGENPOSITION

Kriterien vor der Generierung festzulegen, ist konstruktionsbedingt Exploitation. March (1991) zeigt: Das Verfeinern des Bekannten zahlt sich kurzfristig aus, verdrängt aber die Exploration, die zeigt, was sonst erstrebenswert wäre. Aufsichts-Agenten setzen die heutige Strategie treu durch. Wo lernt die Organisation in diesem Entwurf, dass ihre Regeln falsch sind?

GEGENPOSITION

Verschieden heißt nicht unabhängig. Knight und Leveson (1986) ließen 27 Programme unabhängig nach einer Spezifikation schreiben; sie versagten weit häufiger gemeinsam, als Unabhängigkeit erwarten ließe. Ein Prüfer auf anderem Modell teilt womöglich die blinden Flecken des Generators, weil er die Aufgabe teilt. Lässt sich Unabhängigkeit herstellen oder nur messen?

GEGENPOSITION

Ein Anteil eigener Arbeit reicht womöglich nicht. In der robotischen Chirurgie fand Beane (2019), dass die offizielle Ausbildung versagte, sobald die Technik die Rolle der Auszubildenden am Tisch schrumpfen ließ; Können baute nur eine Minderheit auf, durch regelbrechendes Schattenlernen. Ein budgetierter Übungsanteil bleibt eine verkürzte Lehre. Wer hat dann genug Fehler gesehen, um die Regeln von morgen zu schreiben?

Zweiturteil an einer Stichprobe vom ersten abweicht. Sie sind Diagnosen, keine Zielwerte; wird eine Messgröße zum Ziel, taugt sie nicht mehr (Goodharts Gesetz in der Formulierung von Strathern, 1997). Das Prinzip hinter allen dreien: Nein sagen muss billiger sein als Ja sagen.

6 Fünf Folgerungen für Führungskräfte

1. **Nach der Ebene fragen, nicht nach dem Ob.** Die Ebene benennen, auf der die Fachperson steht, und die, die die Organisation anstrebt.
2. **Die Regeln aufschreiben, bevor weitere Prüfende hinzukommen.** Muss- und Soll-Regeln trennen. Sie wie Code versionieren, im Takt von Recht und Strategie statt der Modellversionen, und Widerspruch zulassen.
3. **Die Rolle der Fachperson neu gestalten, nicht aushöhlen.** Stellenprofil, Anreize und Zeitbudget mit der Arbeit verschieben.
4. **Die Prüfer prüfen.** Stichproben ziehen, Fehler einstreuen, Ablehnungen erfassen. Eine Freigabe, die nie Nein sagt, ist keine Kontrolle, auf keiner Ebene. Wo die KI-Verordnung der EU auf Hochrisikosysteme anwendbar ist, verlangt Artikel 14 wirksame menschliche Aufsicht (Europäische Union, 2024); dass der Mensch auf Ebene 4 wirksam ist, muss gezeigt werden, nicht unterstellt.
5. **Einen Weg zum Aufbau von Expertise offenhalten.** Wer morgen Regeln schreibt, lernt das in der Arbeit von heute.

7 Grenzen

Die Spalte „Morgen“ ist eine Hypothese; die vier Ebenen, die zwei Regelsets und die Rollenleiter sind ein Vorschlag, kein Befund. Einzelfreigabe und Frontloading stammen aus dem Vertrieb von Unternehmenssoftware (Hötter, 2026) und sind anderswo noch nicht validiert. Wie viel einer Strategie sich als prüfbare Soll-Regeln fassen lässt, ist empirisch offen; die Unabhängigkeit von Aufsichts-Agenten und das Erkennen von Drift ohne annotierte Daten sind offene Forschungsprobleme. Nicht jede Aufgabe wird wandern.

Dank. Wir danken Andrew J. Turner, Tim Hourigan, Marilyne Labasque (Begriff „expert in the loop“), Carlos O. Fuentes, Laurent Herbillon und allen weiteren Kommentierenden auf LinkedIn, deren Einwände Abschnitt 4 geprägt haben.

Literatur

- Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint 2212.08073.
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6):775–779.
- Beane, M. (2019). Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, 64(1):87–123.
- Bowman, S. R., Hyun, J., Perez, E., et al. (2022). Measuring progress on scalable oversight for large language models. arXiv preprint 2211.03540.
- Dawes, R. M., Faust, D., und Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.
- Dietvorst, B. J., Simmons, J. P., und Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.
- Europäische Union (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Amtsblatt der Europäischen Union, ABL L, 2024/1689, 12. Juli 2024.
- Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45:105681.
- Hötter, J. (2026). Verification asymmetry in enterprise SaaS sales: Human–AI interface design patterns for AI-based workflows. Masterarbeit, Hasso Plattner Institute, University of Potsdam. Betreut von F. Uebersnickel und T. Wuttke.
- Knight, J. C. und Leveson, N. G. (1986). An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1):96–109.
- March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87.
- Nisbett, R. E. und Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259.
- Parasuraman, R. und Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.
- Skitka, L. J., Mosier, K. L., und Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006.
- Strathern, M. (1997). „Improving Ratings“: Audit in the British university system. *European Review*, 5(3):305–321.
- Uebersnickel, F. (2026). Jede Aufgabe lässt sich mit KI unterstützen. Arbeitspapier, Allgemeine Ausgabe, Oktober 2026, https://uebersnickel.com/papers/Uebersnickel_2026_Jede-Aufgabe-laesst-sich-mit-KI-unterstuetzen_DE.pdf.
- Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: *NeurIPS 2023, Datasets and Benchmarks Track*.
- Uebersnickel · Der Mensch setzt die Regeln · Arbeitspapier · Allgemeine Ausgabe

GEGENPOSITION

Stichproben kalibrieren Vertrauen womöglich in die falsche Richtung. Dietvorst et al. (2015) fanden, dass Menschen das Vertrauen in einen Algorithmus schneller verlieren als in einen Menschen, nachdem beide denselben Fehler gemacht haben, und dann den unterlegenen Menschen vorziehen. Wer die Prüfer stichprobenartig prüft, wird Fehler sehen; die Folge mag Abkehr sein statt Drift. Baut das Prüfen der Prüfer Vertrauen auf, oder zehrt es Vertrauen auf?

GEGENPOSITION

Ein Mensch auf Ebene 4 kann ein Schutzschild sein statt eine Kontrolle. Nach Durchsicht von 41 Vorschriften, die menschliche Aufsicht über staatliche Algorithmen verlangen, schließt Green (2022), die Evidenz spreche dafür, dass Menschen die verlangte Aufsicht nicht leisten können, und solche Vorgaben legitimierten fehlerhafte Systeme, ohne sie zu beheben. Wer trägt die Verantwortung, wenn die Regeln richtig waren und die Prüfer falsch lagen?