

Jede Aufgabe lässt sich mit KI unterstützen

Die Frage ist, wie der Mensch die Führung behält und wie viel Aufsicht das verlangt

Falk Uebernickel

Universität St.Gallen (HSG)

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

Zusammenfassung. In der Debatte über KI in der Arbeitswelt wird häufig gefragt, *ob* eine Aufgabe überhaupt mit KI bearbeitet werden sollte. Dieses Arbeitspapier vertritt eine andere Position: Kaum ein Bereich der Wissensarbeit oder des operativen Geschäfts bleibt, den KI nicht unterstützen kann. Aufgaben unterscheiden sich darin, *wie* KI sie unterstützt und wie viel menschliche Aufsicht diese Form der Unterstützung enthält. Aufbauend auf der Asymmetrie zwischen Erzeugen und Prüfen eines Ergebnisses (Cobbe et al., 2021; Wei, 2025) und auf Gestaltungsmustern aus einer aktuellen Masterarbeit (Hötter, 2026) beschreiben wir eine Matrix aus Generierungs- und Verifikationsaufwand. Jeder Quadrant ist eine Form der KI-Unterstützung mit eigener, eingebauter Aufsicht: Autonom mit Rücknahmerecht, Delegieren mit Einzelfreigabe, Evidenzgeführt über die Zeit und Ko-Kreation mit Freigabe-Gates. Was außerhalb der Matrix bleibt, dem Menschen vorbehalten, ist keine technische Grenze, sondern eine bewusste Entscheidung. Wir veranschaulichen sie mit typischen Aufgaben aus Vertrieb und operativem Geschäft, leiten fünf Folgerungen für Führungskräfte ab und zeigen, wo ein Einstieg sinnvoll ist. Die Beispiele sind illustrativ.

1 Vom Ob zum Wie

Viele KI-Rahmenwerke setzen beim Risiko an: Je höher das Risiko, desto mehr menschliche Beteiligung. Führungskräfte ziehen ihre Linien ähnlich. Strategie, Personalentscheidungen, Kundenbeziehungen und Sicherheit bleiben bei ihnen; Berichtswesen, Terminplanung und Routineentscheidungen ließen sich abgeben. Solche Linien folgen meist der Frage: *Kann KI das, oder nicht?*

Unsere Position ist, dass diese Frage weitgehend beantwortet ist. Nahezu jede Aufgabe lässt sich heute mit KI unterstützen, vom Lieferantendossier bis zur Vorbereitung einer Preisentscheidung. Die nützliche Frage lautet, *wie* KI eine Aufgabe unterstützen soll und wie der Mensch dabei die Führung behält. Eine Risikoklasse beantwortet das nicht: Sie sagt, wie viel menschliche Beteiligung eine Aufgabe braucht, nicht, welcher Art sie sein muss. Man nehme zwei Aufgaben aus dem B2B-Vertrieb, bei denen ähnlich viel auf dem Spiel steht. Eine falsche Antwort in einem Sicherheitsfragebogen kann den Auftrag kosten. Dennoch ist der Fragebogen mühsam zu erstellen, aber schnell zu prüfen, denn jede Antwort verweist auf eine Richtlinie, ein Zertifikat oder eine Vertragsklausel. Ein ROI-Versprechen ist in Minuten geschrieben; ob es hält, zeigt sich erst Monate nach der Einführung. KI kann beide Aufgaben unterstützen. Sie sollte es auf sehr unterschiedliche Weise tun.

2 Zwei Aufwände, vier Formen der KI-Unterstützung

Die Matrix beruht auf einer Beobachtung: Bei vielen Aufgaben ist es weit günstiger, eine Antwort zu prüfen, als sie zu erzeugen. Die Forschung zum maschinellen Lernen nutzt diese Asymmetrie, wenn ein eigener Verifizierer generierte Lösungen bewertet (Cobbe et al., 2021). Wei (2025) formuliert sie als allgemeine Regel, das Gesetz des Verifizierers (*verifier's law*): Wie leicht sich KI für eine Aufgabe trainieren lässt, wächst mit deren Prüfbarkeit. Für Organisationen hat die Asymmetrie eine zweite Bedeutung. Macht KI das Erzeugen billig, entscheidet sich die Form der Unterstützung daran, wie sich das Ergebnis prüfen

GEGENPOSITION

Risiko ist keine Eigenschaft einzelner Ergebnisse. Perrow (1984) verortete die schwersten Unfälle in interaktiv komplexen, eng gekoppelten Systemen, in denen kleine, je für sich harmlose Störungen sich unvorhergesehen verbinden. Eine Verifikationsachse bewertet jede Aufgabe für sich. Sieht eine Matrix je Aufgabe das Risiko zwischen den Aufgaben?

GEGENPOSITION

Die Unterstützung ist womöglich dünner, als sie aussieht. Autor (2015) argumentiert, dass die am schwersten automatisierbaren Aufgaben Flexibilität, Urteilkraft und gesunden Menschenverstand verlangen, die wir nur implizit beherrschen: Polanyis Paradox, wir wissen mehr, als wir sagen können. Wo sich das Schwierige nicht artikulieren lässt, unterstützt KI nur den Papierkram. Ist jede Aufgabe unterstützbar oder nur ihr beschreibbarer Teil?

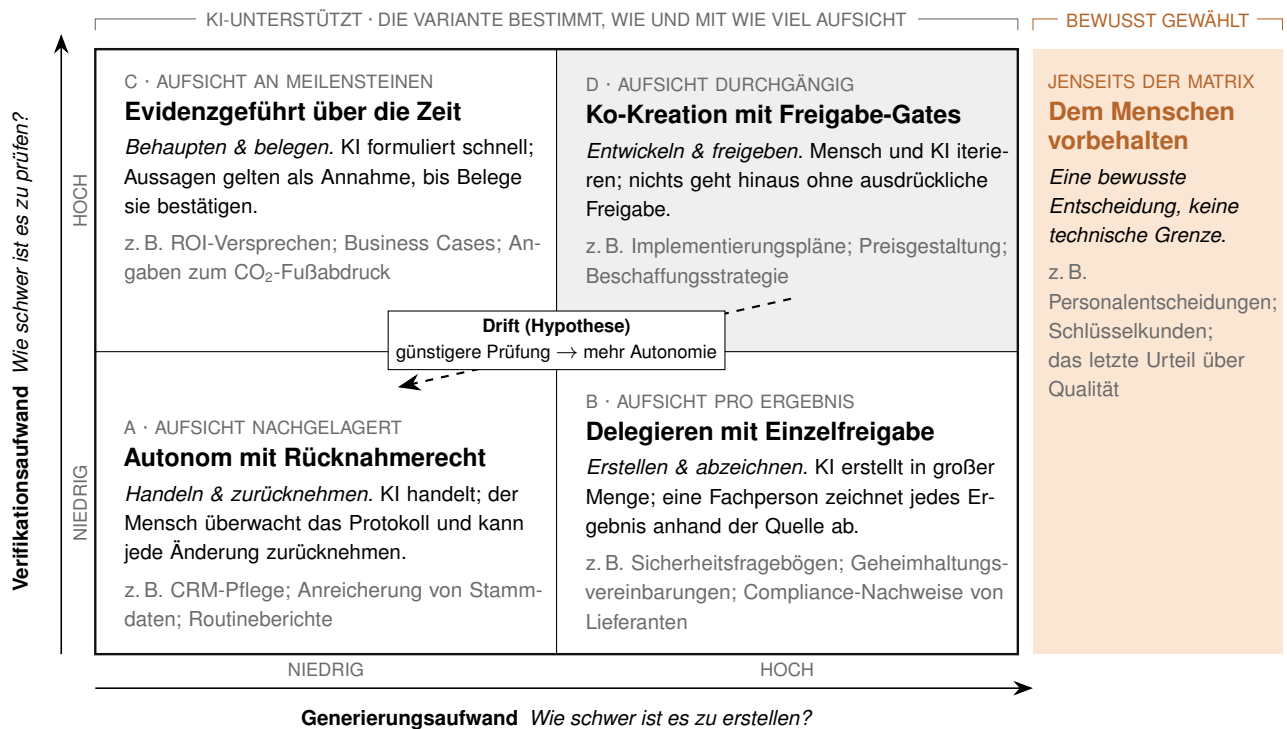


Abbildung 1. Vier Formen der KI-Unterstützung mit Beispielaufgaben aus Vertrieb und operativem Geschäft. Jedes Feld ist KI-unterstützt; was sich unterscheidet, ist die Form der Unterstützung und die darin angelegte Aufsicht. Die Einordnung hängt davon ab, wie eine Organisation eine Aufgabe ausgestaltet, und ist als Diskussionsgrundlage gedacht. Nach Hötter (2026); Achsen nach Wei (2025).

- **Generierungsaufwand** (horizontal): Wie schwer ist es, das Ergebnis zu erstellen? Etwa das erforderliche Fachwissen, die Zahl der Quellen, der Umfang des Kontexts.
- **Verifikationsaufwand** (vertikal): Wie schwer ist es, das Ergebnis zu prüfen? Etwa die Prüfdauer, die Zahl der beteiligten Rollen, die Kosten einer Korrektur.

A • Autonom mit Rücknahmerecht (niedrige Generierung, niedrige Verifikation). Die KI handelt; der Mensch überwacht das Protokoll und behält das Recht, zurückzunehmen. Das eignet sich für umkehrbare Arbeit in großer Zahl, etwa das Aktualisieren von CRM-Datensätzen, das Anreichern von Stammdaten oder das Fortschreiben eines Lieferstatus. Es scheitert, wo die Rücknahme fehlt. Eine Kundenantwort lässt sich, einmal gegeben, nicht zurückholen: In *Moffatt v. Air Canada* (2024 BCCRT 149) haftete eine Fluggesellschaft für die Auskunft des Chatbots auf ihrer Website. Irreversibilität schiebt eine Aufgabe auf der Verifikationsachse nach oben, unabhängig von ihrem Generierungsaufwand.

B · Delegieren mit Einzelfreigabe (hohe Generierung, niedrige Verifikation). Die KI erstellt in großer Menge; eine Fachperson zeichnet jedes Ergebnis anhand seiner Quelle ab. Typische Fälle sind die Nachweisdokumente, von denen der Geschäftsalltag ohnehin lebt: Sicherheitsfragebögen und Geheimhaltungsvereinbarungen im Vertrieb, Compliance-Nachweise von Lieferanten, regulatorische Unterlagen. Hier nimmt die KI den größten Teil der Erstellungsarbeit ab, und das Risiko verlagert sich auf die prüfende Person. Wer überwiegend korrekte Ergebnisse freigibt, gewöhnt sich an, alles freizugeben (Parasuraman und Riley, 1997; Skitka et al., 1999), und seltene Fehler werden weit häufiger übersehen als

häufige (Wolfe et al., 2005). Eine Freigabe, die nie Nein sagt, ist keine Kontrolle.

C • Evidenzgeführt über die Zeit (niedrige Generierung, hohe Verifikation). Behauptungen sind schnell geschrieben und langsam zu bestätigen: der Ertrag eines neuen Werkzeugs, eine Prognose, eine Verringerung des CO₂-Fußabdrucks. Die KI entwirft sie in Minuten. Die Aufsicht liegt in dem, was danach geschieht: Jede Aussage bleibt als Annahme gekennzeichnet, erhält eine verantwortliche Person und einen Termin für den Beleg und wird von der zuständigen Rolle abgenommen, sobald der Beleg vorliegt. Das Scheitern geschieht leise. Annahmen verfestigen sich zu Tatsachen, weil niemand sie wieder aufgreift.

D • Ko-Kreation mit Freigabe-Gates (hohe Generierung, hohe Verifikation). Auch strategische und gestalterische Arbeit, etwa ein Preismodell, ein Implementierungsplan oder eine Beschaffungsstrategie bei volatilen Preisen, lässt sich mit KI unterstützen. Hier erarbeiten Mensch und KI das Ergebnis gemeinsam, und ausdrückliche Freigabestufen entscheiden, was nach außen geht. Feldbefunde zeigen, warum diese Gates wichtig sind. In einem Experiment mit Beraterinnen und Beratern erhöhte KI die Qualität bei Aufgaben innerhalb ihrer Leistungsgrenze, senkte aber die Richtigkeit bei einer Aufgabe knapp außerhalb, bei der die Beratungskräfte ihr zu bereitwillig vertrauten (Dell'Acqua et al., 2023).

Dem Menschen vorbehalten: eine bewusste Entscheidung. Lässt sich nahezu jede Aufgabe mit KI unterstützen, dann ist die Arbeit, die rein menschlich bleibt, kein Rest mehr, den die Technik noch nicht erreicht hat. Sie ist eine Entscheidung. Eine Organisation kann festlegen, dass Personalentscheidungen, ihre wichtigsten Kundenbeziehungen und das letzte Qualitätsurteil beim Menschen bleiben, weil der menschliche Akt selbst den Wert trägt: Handwerk, Verantwortung, Vertrauen. Eine solche Entscheidung sollte benannt, begründet und regelmäßig überprüft werden, statt als stilles Zögern stehen zu bleiben.

3 Drift: Human-in-the-Loop als Übergang

Die Matrix ist nicht statisch. Je günstiger die Prüfung wird, durch verknüpfte Quellen, strukturierte Ausgaben, automatisierte Tests oder regelbasierte Kontrollen, desto weiter wandern Aufgaben nach unten links, Richtung „Autonom mit Rücknahmerecht“. In der Softwareentwicklung, wo sich jede Änderung gegen eine automatisierte Testsuite prüfen lässt, lösen Modelle einen wachsenden Anteil realer Aufgaben (Jimenez et al., 2024). Wir betrachten die Freigabe jedes einzelnen Ergebnisses deshalb als Übergangstechnologie, als Hypothese, nicht als Befund. Die Folge: Die Rolle des Menschen verändert sich, statt zu verschwinden. Sie verschiebt sich vom Prüfen jedes einzelnen Ergebnisses zur Festlegung dessen, was als richtig gilt, und zur Prüfung der Prüfer, des Forschungsprogramms, das unter dem Begriff *Scalable Oversight* (skalierbare Aufsicht) firmiert (Bowman et al., 2022).

4 Fünf Folgerungen für Führungskräfte

1. **Nach dem Wie fragen, nicht nach dem Ob.** Davon ausgehen, dass sich jeder Prozess mit KI unterstützen lässt. Bevor ein Anwendungsfall als „zu riskant“ zurückgestellt wird, ist zu fragen, wie sich sein Ergebnis prüfen ließe; die Antwort ergibt die Variante.
2. **Bewusst festlegen, was dem Menschen vorbehalten bleibt.** Je Funktion vereinbaren, welche Tätigkeiten menschlich bleiben, und festhalten, warum. Eine ausdrückliche, regelmäßig überprüfte Entscheidung schützt Handwerk und Kundenbeziehungen besser als ein diffuses Zögern.
3. **Die Zeit der Fachleute als Engpass budgetieren.** Das Erzeugen wird billig, das Prüfen nicht. Jeder Anwendungsfall der Variante „Delegieren mit Einzelfreigabe“ verbraucht die Stunden jener Menschen, auf deren Urteil die Organisation angewiesen ist.
4. **Messen, was die Aufsicht findet.** Erfassen, was die Prüfenden ändern und warum, und gelegentlich bekannte Fehler einstreuen, um zu sehen, ob sie gefunden werden.

GEGENPOSITION

Wer eine Aussage verantwortet, ist selten ihr bester Prüfer. In 258 Verkehrsinfrastrukturprojekten zeigten Flyvbjerg et al. (2002), dass sich die Kostenunterschätzung besser durch strategische Falschdarstellung als durch Irrtum erklären lässt. Wer hält den Beleg, wenn der Urheber den Termin hält?

GEGENPOSITION

Ko-Kreation bleibt womöglich nicht Ko-Kreation. Noy und Zhang (2023) fanden, dass ChatGPT die Schreibproduktivität von Berufstätigen deutlich erhöhte und, in der Lesart der Autoren, den Arbeitseinsatz eher ersetzte als Fähigkeiten ergänzte; in der Working-Paper-Fassung reichten die meisten Teilnehmenden, die ChatGPT nutzten, dessen ersten Entwurf unverändert ein. Ist gemeinsames Entwickeln eine stabile Form oder ein Abgleiten in Delegation?

GEGENPOSITION

Was menschlich bleibt, entscheidet womöglich der Kunde, nicht die Firma. Longoni et al. (2019) fanden, dass Verbraucherinnen und Verbraucher Gesundheitsleistungen eines KI-Anbieters seltener nutzten, aus Sorge, KI könne ihre Umstände nicht berücksichtigen (uniqueness neglect); der Widerstand ließ nach, wenn KI einen menschlichen Anbieter unterstützte statt ersetzte. Wer entscheidet, was menschlich bleibt?

GEGENPOSITION

Günstigere Prüfungen prüfen, wofür sie geschrieben wurden. In einer Nutzerstudie fanden Perry et al. (2023), dass Teilnehmende mit KI-Assistent deutlich unsichereren Code schrieben als solche ohne und ihren Code eher für sicher hielten. Code kann seine Tests bestehen und dennoch unsicher sein. Macht günstigere Verifikation Autonomie sicherer, oder lässt sie Drift nur sicher erscheinen?

Eine Freigabequote nahe 100 % ist ein Warnsignal, kein Erfolg.

5. **Die Regeln aufschreiben.** Trennen, was nie erlaubt ist (Recht, Compliance, Branchenstandards), von dem, was die Führung will (Unternehmensstandards, Servicestil). Schriftliche, versionierte Regeln sind die Voraussetzung dafür, dass Aufsicht skaliert. Wo die KI-Verordnung der EU auf Hochrisikosysteme anwendbar ist, verlangt ihr Artikel 14 eine wirksame menschliche Aufsicht (Europäische Union, 2024).

5 Wo beginnen

Wir schlagen vor, ein Portfolio von KI-Initiativen systematisch aufzubauen statt Fall für Fall. Ein erster Schritt kann ein kleines Strategiprojekt sein, das festlegt, wo begonnen wird, jeden Kandidaten in der Matrix verortet und für jede Funktion die dem Menschen vorbehaltenen Tätigkeiten benennt. Im Portfolio sollte die *Augmentierung* Vorrang haben: Anwendungsfälle, in denen KI die Arbeit der Menschen stärkt, während Urteil und Verantwortung bei ihnen bleiben, wie in den Varianten „Delegieren mit Einzelfreigabe“, „Evidenzgeführt über die Zeit“ und „Ko-Kreation mit Freigabe-Gates“. Wer dort beginnt, baut Prüfroutinen, schriftliche Regeln und Messgrößen für das auf, was die Aufsicht findet (Abschnitt 4). Von ihnen hängt jeder Schritt Richtung „Autonom mit Rücknahmerecht“ ab.

6 Grenzen

Dass sich jede Aufgabe mit KI unterstützen lässt, ist eine Position, kein empirischer Befund, und Unterstützung bedeutet nicht Automatisierung. Die Matrix wurde aus dem Vertrieb von Unternehmenssoftware abgeleitet (Hötter, 2026) und in Bereichen wie Fertigung oder Dienstleistungen noch nicht validiert. Die Beispiele sind illustrativ und beschreiben keine bestimmte Organisation. Ein Portfolio wie das vorgeschlagene würde zugleich die Matrix prüfen: einige Anwendungsfälle über zwei bis drei Quartale verorten, betreiben und messen.

GEGENPOSITION

Augmentierung hat ihren eigenen Preis. Über 106 Experimente hinweg fanden Vaccaro et al. (2024), dass Mensch-KI-Kombinationen im Mittel schlechter abschnitten als der bessere von Mensch oder KI allein: Gewinne, wo Menschen der KI überlegen waren, Verluste, wo die KI den Menschen überlegen war. Welche Aufgaben der Variante „Delegieren mit Einzelfreigabe“ erledigte die KI allein besser?

Literatur

- Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30.
- Bowman, S. R., Hyun, J., Perez, E., et al. (2022). Measuring progress on scalable oversight for large language models. arXiv preprint 2211.03540.
- Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). Training verifiers to solve math word problems. arXiv preprint 2110.14168.
- Dell’Acqua, F., McFowland III, E., Mollick, E. R., et al. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School.
- 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.
- Flyvbjerg, B., Holm, M. S., und Buhl, S. (2002). Underestimating costs in public works projects: Error or lie? *Journal of the American Planning Association*, 68(3):279–295.
- 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. Uebernickel und T. Wuttke.
- Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., und Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? In: *International Conference on Learning Representations (ICLR)*.
- Kabir, S., Udo-Imeh, D. N., Kou, B., und Zhang, T. (2024). Is Stack Overflow obsolete? An empirical study of the characteristics of ChatGPT answers to Stack Overflow questions. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24)*, S. Article 935, Honolulu. ACM.
- Longoni, C., Bonezzi, A., und Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650.
- Noy, S. und Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.
- Parasuraman, R. und Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.
- Perrow, C. (1984). *Normal Accidents: Living with High-Risk Technologies*. Basic Books, New York.
- Perry, N., Srivastava, M., Kumar, D., und Boneh, D. (2023). Do users write more insecure code with AI assistants? In: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS ’23)*, S. 2785–2799, Copenhagen. ACM.
- 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.
- Vaccaro, M., Almaatouq, A., und Malone, T. W. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303.
- Wei, J. (2025). Asymmetry of verification and verifier’s law. Blogbeitrag, <https://www.jasonwei.net>.
- Wolfe, J. M., Horowitz, T. S., und Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, 435(7041):439–440.