Human-in-the-Loop: Warum menschliche Kontrolle bei KI schnell zur Illusion wird

KI wird in Unternehmen zunehmend fester Bestandteil der Wertschöpfung (z. B. in der Softwareentwicklung, im Service, im Personalbereich). Für Entscheidende entsteht dadurch ein Dilemma: Einerseits müssen die Produktivitätspotenziale realisiert werden, andererseits gilt es, Risiken und oft noch unklare Haftungsfragen wie bspw. bei Fehlentscheidungen der KI zu managen (siehe bitkom: Wie steht es um Künstliche Intelligenz in Deutschland, 2025).

Die aktuelle Welle von Agentic AI mit der Integration von KI-basierten Automatisierungen in die eigenen Geschäftsprozesse verschärft dieses Dilemma. KI ist nicht mehr nur ein reaktiver Chatbot, sondern recherchiert, plant, schreibt, macht und tut. Die Integrationen reichen von isolierten Workflow-Engines und Office-Produkten bis hin zu kritischen CRM- und ERP-Systemen, in denen KI in operative Abläufe eingebunden wird. Dadurch verschieben sich jedoch auch die Risiken, denn KI erzeugt nun nicht mehr „nur“ fehlerhafte Antworten, sondern falsche Ausgaben und Entscheidungen werden durch die Prozesse hindurch propagiert mit entsprechenden Schäden. Daher ist es selbstverständlich, dass Unternehmen nach Sicherheitsmechanismen suchen und schnell Lösungen brauchen. Einer dieser häufig genutzten Mechanismen ist **Human-in-the-Loop.**

Human-in-the-Loop verkörpert die Standardantwort auf Automatisierungsrisiken: „Am Ende schaut noch ein Mensch drauf.“ Es wirkt attraktiv, weil es Compliance, Datenschutz und den Betriebsrat beruhigt und leicht in der Umsetzung ist. Es braucht keinen großen Aufwand, sondern vorhandene Experten werden mit dieser Aufgabe betraut. Es signalisiert Verantwortlichkeit gegenüber Kunden und Führungskräften und zuletzt verspricht es einen KI-Einsatz, ohne sich vollständig auf die Maschine verlassen zu müssen.

Was bedeutet „Human-in-the-Loop“ (HITL)?

Bei Human-in-the-Loop wird der Mensch aktiv in einen (teil-)automatisierten Entscheidungsprozess eingebunden. Der Mensch prüft, korrigiert oder bestätigt KI-Vorschläge, bevor sie wirksam werden. Das Ziel ist die Kombination der Effizienz der Automatisierung mit menschlichem Urteilsvermögen, Kontextwissen und ethischer Abwägung. So sehr HITL auch vielversprechend klingt, setzt es voraus, dass Menschen KI-Aufgaben überhaupt zuverlässig prüfen **können, wollen und dürfen.**

Die Human-in-the-Loop Schwachstelle: Der Mensch

„Die KI macht die Arbeit – der Mensch bestätigt.“ So stellen sich viele Organisationen Human-in-the-Loop vor. Ein Sicherheitsmechanismus, der KI-Entscheidungen absichert: Die Maschine liefert Vorschläge, der Mensch prüft und entscheidet mit aller vorhanden Erfahrung und Wissen. In der Praxis funktioniert das selten so reibungslos. Oft überschreiben Menschen korrekte KI-Empfehlungen, folgen den falschen KI-Empfehlungen und sie erkennen nicht, wann sie der KI vertrauen sollten und wann nicht. Daher darf **Human-in-the-Loop nicht als pauschale Lösung gelten**, sondern muss für den konkreten Anwendungsfall gestaltet werden. Andernfalls sind Menschen zwar formell beteiligt, aber es gibt keine faktisch wirksame Kontrolle. Als prominentes Beispiel wie HITL schief gehen kann möchten wir den Starbucks-„Tank Day“-Skandal in Südkorea nennen, da dort sichtbar wird wie fehlende menschliche Kontrolle bei KI-gestütztem Marketing zu gravierenden Reputationsrisiken führen kann (The Guardian: How a Starbucks marketing stunt spiralled into mass boycotts in South Korea).

Der verbreitete Irrtum lautet: Sobald ein Mensch beteiligt ist, wird die Entscheidung automatisch besser. Die Forschung zeigt jedoch sogar ein differenzierteres Bild. In vielen Studien schneiden **hybride Entscheidungen aus Mensch und KI schlechter ab** als reine KI-Entscheidungen.

Menschen im HITL-Prozess sind keine neutralen, rationalen „Fehlerfilter". Sie unterliegen kognitiven Verzerrungen, stehen unter Zeit- und Produktivitätsdruck und verfügen häufig weder über eine ausreichende Erklärbarkeit der Systeme noch über klare Verantwortlichkeiten. Verhaltensökonomische Forschung zeigt zudem, dass Menschen systematisch irrational handeln – selbst dann, wenn sie sich dieser Verzerrungen bewusst sind.

Der Grund liegt in den bereits erwähnten Verzerrungen:

- Menschen **überschreiben korrekte KI-Empfehlungen**
- sie **folgen falschen KI-Empfehlungen**
- und sie erkennen oft nicht, **wann Vertrauen in die KI sinnvoll ist.**

Die Konsequenz ist jedoch nicht, Human-in-the-Loop abzuschaffen oder nicht zu nutzen.

Im Gegenteil: Gerade weil Menschen bestimmten psychologischen Mustern folgen, müssen Organisationen **HITL bewusst gestalten**. Andernfalls entsteht schnell eine Illusion von Kontrolle: Das Unternehmen glaubt, ein Risiko entschärft zu haben, obwohl es dieses lediglich verschoben hat.

Im Folgenden betrachten wir anhand von Erkenntnissen aus der Verhaltensökonomie, warum Human-in-the-Loop häufig am Menschen scheitert – und welche Strukturen Organisationen benötigen, damit menschliche Aufsicht tatsächlich funktioniert. Am Ende des Artikels gehen wir auch auf Lösungsansätze und konkrete Maßnahmen ein, die auch Teil der umfassenden KI-Beratung von viadee spark sind.

Prüfen, winken und lächeln

Viele Organisationen senden widersprüchliche Signale. Auf der einen Seite sollen Mitarbeitende KI nutzen, um die Produktivität zu steigern. Auf der anderen Seite sollen sie KI-Ergebnisse gründlich prüfen. Diese beiden Ziele stehen häufig im direkten Konflikt. Wer sich gegen eine KI-Empfehlung entscheidet, erzeugt zusätzliche Arbeit: mehr Recherche, mehr Begründung, möglicherweise Diskussionen mit Kolleg:innen oder Führungskräften. **In der Praxis bedeutet das oft: Entscheidungen werden einfach durchgewunken, weil „der Prozess halt so ist".**

HITL und der Automation Bias

Dazu kommt ein zentrales Paradoxon moderner Automatisierung: der Automation Bias.

Wenn KI-Systeme in den meisten Fällen korrekt arbeiten und nur selten Fehler machen, wird menschliche Kontrolle paradoxerweise schwieriger. Die Prüfung wirkt schnell wie eine Formalität. In der Praxis folgt daraus ein rationales Verhalten: Menschen entscheiden sich im Zweifel für die KI-Empfehlung, um Aufwand zu vermeiden und Produktivitätsziele zu erreichen, denn **Menschen bleiben nicht dauerhaft aufmerksam, wenn fast immer alles richtig ist**.

Der Effekt wurde bereits früh beschrieben, etwa in Lisanne Bainbridges Konzept der „Ironies of Automation": Je zuverlässiger ein automatisiertes System wird, desto schwieriger wird es für Menschen, es sinnvoll zu überwachen.

In solchen Situationen entwickelt sich die KI-Empfehlung faktisch zum Status quo. Abweichungen davon erfordern zusätzliche Begründungen und kosten Zeit. Dieser Effekt wird verstärkt, wenn die KI-Ausgabe als „Empfehlung“ oder Standardentscheidung präsentiert wird. Durch dieses Framing wird die KI-Option überproportional häufig gewählt.

Wie erklärbar sind Entscheidungen noch?

Hinzu kommt ein weiteres Problem: fehlende Erklärbarkeit.

Die meisten KI-Modelle liefern plausible Ergebnisse, ohne transparent zu machen, wie diese zustande gekommen sind. Für eine echte Kontrolle reicht es jedoch nicht, wenn ein Ergebnis „vernünftig klingt“. **Wenn Prüfer:innen nicht nachvollziehen können, auf welchen Annahmen oder Daten ein Ergebnis basiert, bleibt nur eine oberflächliche Plausibilitätsprüfung.**

Auch soziale Normen spielen eine Rolle: Wenn KI unternehmensweit als Effizienzprogramm eingeführt wird, entsteht schnell der Eindruck: Alle arbeiten mit KI – wir verlassen uns auf das System.

Wann prüft der Mensch, wann nickt er einfach ab?

Das Resultat ist eine Verantwortungsdiffusion. Am Ende hat scheinbar „die KI entschieden“, nicht die handelnde Person. Der Mensch wird damit zum Abnicker statt zum Prüfer.

Organisationen müssen daher Prozesse gestalten, in denen Menschen nicht nur formal beteiligt sind, sondern auch tatsächlich die Chance haben, Fehler zu erkennen. Regulatorische Anforderungen des EU AI Act betonen genau diese Aspekte: Transparenz, Risikomanagement und wirksame menschliche Aufsicht.

Der Mensch löst Probleme “ausreichend gut”

Wenn Menschen tatsächlich gründlich prüfen sollen, entsteht ein erheblicher Aufwand. Eine echte Kontrolle bedeutet mehr als nur ein kurzes Überfliegen des Ergebnisses. Prüfer:innen müssten:

- Fakten und Quellen überprüfen
- die zugrunde liegenden Daten validieren
- Annahmen und Zwischenschritte nachvollziehen
- fachliche, rechtliche und technische Aspekte berücksichtigen

In vielen Fällen ist diese Prüfung **fast so aufwendig wie die ursprüngliche Bearbeitung**.

Damit entsteht das grundlegende Dilemma:

- Entweder Organisationen realisieren **Produktivitätsgewinne**, verzichten aber auf gründliche Kontrolle.
- Oder sie prüfen sorgfältig – und verlieren einen großen Teil des Effizienzvorteils.

Neben den reinen Arbeitsaufwänden entsteht psychologischer Druck, **da der Mensch die Verantwortung für ein Ergebnis trägt, dessen Entstehung er nicht vollständig kontrollieren kann**. Fehler werden ihm zugerechnet, obwohl System, Modell, Daten und Prozessgestaltung die Ursache sein können.

Mit der zunehmenden Verbreitung **agentischer Systeme** verschärft sich dieses Problem weiter. Wenn KI tausende Tickets bearbeitet, Millionen Transaktionen bewertet oder Entscheidungen in Echtzeit trifft, kann kein Mensch mehr jede Entscheidung prüfen.

Unter Zeitdruck greifen Menschen dann auf Heuristiken zurück – einfache Entscheidungsregeln wie: *„Wenn die KI nix anzeigt, wird es schon passen.“* Diese Mechanismen wurden bereits von **Tversky und Kahneman** beschrieben: Menschen reduzieren Komplexität durch mentale Abkürzungen.

Doch genau diese Heuristiken können in hochautomatisierten Systemen riskant werden.

Wie gestalten wir menschliche Aufsicht so, dass sie tatsächlich wirksam ist?

Das Dilemma mit HITL ist nicht nur theoretisch, sondern bereits in der Praxis zu sehen wie beim Eingangs benannten Beispiel zu sehen ist. Wenn Human in the Loop falsch gestaltet wird und ohne entsprechende Befähigung stattfindet, entstehen konkrete strategische, wirtschaftliche und regulatorische Risiken. Um auf diese Risiken einzugehen, gibt es drei Optionen für Unternehmen:

Option 1: Verzicht auf KI

Die naive Konsequenz des Dilemmas ist, KI nicht zu nutzen. Wenn menschliche Kontrolle zu teuer ist, man das Risiko des KI-Einsatzes aber nicht eingehen will, kann man schlussfolgern, KI nicht einzusetzen. Diese Schlussfolgerung führt zu einer paradoxen Situation. Eine technisch überlegene Lösung wird nicht eingesetzt, nicht weil sie schlecht ist, sondern weil der menschliche Kontrollmechanismus ökonomisch nicht funktioniert (siehe: The Human-AI Contracting Paradox). Ein Unternehmen würde bei dieser Option also auf einen wichtigen Wettbewerbsvorteil verzichten.

Option 2: Verzicht auf Kontrolle

Da KI-Systeme ohne menschliche Interaktion häufig performanter sind als HITL oder hybride Systeme, könnte man schlussfolgern, direkt auf HITL und menschliche Kontrolle zu verzichten. In diesem Fall sollte man vor einem Rollout Vergleichstests durchführen, um zu prüfen, inwieweit sich Performance und Fehlerraten zwischen einem reinen KI- und einem HITL-System unterscheiden. Wie steht es um Bearbeitungszeiten, Fehlertypen, Eskalationsquoten, Nacharbeitskosten oder Auditierbarkeit?

Leistet man diese wichtigen empirischen Vorarbeiten nicht, riskiert man den Kontrollverlust gegenüber KI-Agenten sowie ein hohes Kostenrisiko. Je agentischer KI-Systeme werden, desto größer wird das Risiko von Folgeschäden z. B. falsche Bestellungen, fehlerhafte Kundenkommunikation oder unnötige Cloud- oder API-Kosten. Wenn nicht jede Ausgabe menschlich geprüft wird, müssen klare Kriterien für die Prüfung sowie Stichproben mit Schwellenwerten und Monitoring etabliert werden, um das Risiko zu steuern. Technische und organisatorische Kontrollen sollten Budgetlimits, Rechte- und Rollenkonzepte, Transaktionsgrenzen, Sandbox-Umgebungen, Freigabestufen, Logging und Audit Trails umfassen.

Option 3: Aktiv Bedingungen für Kontrolle schaffen

In hochregulierten Bereichen ist menschliche Aufsicht unverzichtbar. Regulierung wie der EU AI Act verlangt in bestimmten Kontexten sogar menschliche Aufsicht. Ohne Human-in-the-Loop würden Unternehmen riskieren, kritische Entscheidungen vollständig zu automatisieren.

Allerdings ist ein Mensch im Prozess keine hinreichende Kontrolle.

Human-in-the-Loop funktioniert nur, wenn ein Unternehmen die dafür notwendigen Bedingungen schafft. Es braucht:

- klare Verantwortlichkeiten
- ausreichende Ressourcen
- echte Entscheidungsbefugnisse und Eskalationswege sowie organisatorische und technische Kontrollmechanismen.

Ohne diese Bedingungen ist Human-in-the-Loop nicht Governance, sondern performatives Theater.

Was zu tun ist, um HITL wirksam zum Einsatz zu bringen?

Statt pauschal Human in the Loop zu fordern, müssen Organisationen konkret definieren:

- **Welche Entscheidungen** prüft der Mensch?
- **Wie viele Fälle** werden kontrolliert – alle, Stichproben oder nur Hochrisikofälle?
- **Wie schnell** muss eine menschliche Reaktion erfolgen?

Ohne diese Klarheit bleibt HITL ein vages Versprechen.

Wir stellen 5 Maßnahmen vor, um HITL erfolgreich einsetzen zu können:

Maßnahme 1: Zeit für Prüfung explizit einräumen und planen

Die Prüfung von KI-Output muss als Arbeitsaufwand geplant werden, nicht als unsichtbare Zusatzaufgabe für Mitarbeitende. Konkret bedeutet dies, HITL bei der Bewertung von KI-Use-Cases sowie in Projektplänen und Schätzungen zu berücksichtigen und die Produktivitätsziele entsprechend anzupassen. Es darf keine Erwartung erzeugt werden, dass KI gleichzeitig massiv Zeit spart und vollständig geprüft wird. Leitfrage: Wenn ein Mensch verantwortlich prüfen soll: hat er realistisch genug Zeit, um *Nein* zu sagen?

Maßnahme 2: Stichproben, Monitoring und risikobasierte Kontrolle

Nicht jeder KI-Output muss vollständig geprüft werden. Die Prüftiefe sollte je Risikoklasse definiert werden.

High-risk: vollständige Prüfung oder Freigabe vor der Aktion, Medium-risk: Stichproben plus Monitoring, Low-risk: automatisierte Kontrolle, z. B. Logging oder ex-post-Review. Das Monitoring sollte Fehlerquoten, Fehlerarten, Korrekturraten durch Menschen, Durchwinkquoten, Eskalationen, Abweichungen zwischen Teams und Kosten der Prüfung messen. Wichtig: Dass Menschen KI-Ausgaben nie korrigieren, ist kein automatischer Qualitätsbeweis, sondern womöglich ein Anzeichen des Survivorship-Biases oder des Automation-Biases sowie fehlender Prüfung.

Maßnahme 3: Konsequenzen und Anreize richtig setzen

Mitarbeitende müssen belohnt werden, wenn sie Fehler entdecken — nicht bestraft, weil sie Prozesse verlangsamen. Hier ist auf Psychologische Sicherheit zu achten, um Nachfragen und Zweifel jeder Art zuzulassen. Konkret sind neben Produktivitätsmetriken auch Qualitätsmetriken wichtig. Gefundene KI-Fehler sollten als Wertbeitrag und nicht als extra Aufwand interpretiert werden. Klare Eskalationspfade und Verantwortlichkeiten sorgen für einen geregelten Ablauf und entsprechende Haftungsfragen. Es sollte signalisiert werden, dass ein *schnelles Durchwinken* nicht erwünscht ist.

Maßnahme 4: Technische Guardrails einbauen

Technische Guardrails sind oft zuverlässiger als nachträgliche menschliche Kontrolle. Der Mensch sollte nicht die einzige Sicherheitsbarriere sein. Besonders bei skalierten Prozessen muss Sicherheit in Architektur und Prozess eingebaut sein. Beispiele umfassen Zugriffsbeschränkungen, Rollen- und Rechtekonzepte, Tool-Permissions für Agenten, Budget- und Transaktionslimits, Freigabeschwellen, Prompt- und Output-Filter, Retrieval nur aus freigegebenen Quellen, automatische Quellenprüfung, Logging und Audit Trails, Kill Switches, Sandbox-Umgebungen, Policy-as-Code.

Maßnahme 5: Fehler akzeptieren und Regressmöglichkeiten

Null Fehler sind kein realistisches Ziel. Für fehlerhafte KI-Ausgaben, die einen Geschäftsprozess durchlaufen, müssen Regressmöglichkeiten etabliert werden, um Fehler rückwirkend zu korrigieren. Die Regressmöglichkeiten können kontigent an den Merkmalen der Fehler ausgerichtet sein, z. B. welche Fehler sind akzeptabel oder kritisch? Wie schnell werden Fehler erkannt und korrigiert? Wer trägt die Kosten? Die letzten Fragen bedürfen bei einem Portfolio an internen und externen KI-Anwendungen einer Klärung, wer für welche Tätigkeit und Ausgabe verantwortlich ist, z. B. mit einer Verantwortlichkeitsmatrix.

Entscheidend ist, dass Unternehmen Fehlerbudgets für KI definieren, mit klaren Schwellenwerten für Reaktionen (z. B. Alles gut, Eskalation, Abschaltung, Anpassung). Diese Pläne und Maßnahmen sind hilfreicher, als *ein Mensch prüft das, also ist es sicher*.

Fazit

Human-in-the-Loop ist keine universelle Lösung, sondern ein **Gestaltungsproblem**. Wenn Organisationen menschliche Aufsicht nur formell einführen, entsteht keine echte Kontrolle.

Wir freuen uns darauf, das Gespräch zur sicheren Nutzung von KI fortzuführen. Vom Kaffee über den Workshop bis hin zur ganzheitlichen Organisationsentwicklung sprecht uns gerne an.

Mehr Infos hier: KI-Beratung: Künstliche Intelligenz wirksam integrieren.