

PRAKTISCHES WERKZEUG · VERANTWORTUNGSVOLLER KI-EINSATZ

Human-in-the-loop

Welcher Mensch validiert, nach welchem Kriterium, mit welchem Vetorecht, zu welcher Eskalation?

Dieses Werkzeug vermeidet die trügerische Sicherheit von «die KI hilft ja nur». Es zwingt dazu, Prozess für Prozess den Moment zu benennen, in dem ein Mensch verantwortlich bleibt: vor einer sensiblen Entscheidung, einer Kundennachricht, einer HR-Massnahme, einer Zahlung, einer Datenänderung oder bei Unsicherheit des Modells.

ANLEITUNG

1

Prozess zerlegen
Eine Zeile = ein Schritt (max. 3 bis 5). Format:
Eingabe → KI schlägt vor → Mensch prüft → gibt frei oder blockiert → Aktion.

2

Kritische Punkte erkennen
Schritte markieren, die Kunden, HR, Finanzen, Verträge, Recht, Reputation oder sensible Daten betreffen.

3

Menschliche Kontrolle benennen
Wer prüft was, nach welchem Kriterium, vor welcher Aktion. Die Person kann blockieren, korrigieren, Rat einholen oder auf manuell zurückgehen — ohne aufwendige Begründung.

4

Entscheiden

● **Aufsicht ausreichend:** interne Nutzung, nicht sensible Daten, geprüfte Ausgabe, geringe Auswirkung.

● **Aufsicht verstärken:** Kunde / Vertrag / Geld / HR → menschliche Freigabe vor der Aktion zwingend.

● **Prozess zu autonom:** kein echtes Veto oder automatisierte Einzelentscheidung → Design überarbeiten, nicht nur ein «Freigabe»-Kästchen hinzufügen.

PROZESSSCHRITT	ROLLE DER KI	MENSCHLICHE KONTROLLE	FREIGABEKRITERIUM	VETO / ESKALATION	VERDIKT
Eingehendes Kundenticket	Klassiert die Anfrage	Support prüft Kategorie + Dringlichkeit	Kategorie stimmig, Dringlichkeit korrekt	Ja — Support-Leitung	● Ausreichend
Entwurf einer Kundenantwort	Schlägt eine Nachricht vor	Mitarbeitende/r prüft vor Versand	Richtigkeit + Ton + kein riskantes Versprechen	Ja — Vorgesetzte/r / Rechtsabteilung	● Zu verstärken
Vorsortierung von Bewerbungen	Fasst zusammen / bewertet Profile	HR vergleicht mit vollständigem Dossier	Keine automatische Ablehnung	Ja — HR-Leitung + DSB	● Zu autonom

SCHNELLE FRAGEN ZUM AUSFÜLLEN DES VERDIKTS

Sensible Entscheidung? Beeinflusst die KI eine Entscheidung über eine Person, einen Kunden, einen Vertrag, eine Zahlung oder ein Recht?

Sensible Daten? Nutzt der Prozess HR-, Kunden-, Finanz-, Vertrags-, medizinische oder vertrauliche Daten?

Kostspieliger Fehler? Kann eine falsche KI-Ausgabe finanziellen, rechtlichen, reputativen oder menschlichen Schaden verursachen?

Kompetenter Mensch? Versteht die freigebende Person das Fachgebiet, die Grenzen des Tools und die Entscheidungskriterien?

Echtes Vetorecht? Kann sie ohne aufwendige Begründung und ohne Sanktion blockieren oder korrigieren?

Klare Eskalation? Ist im Zweifel klar, an wen weitergeleitet wird: Vorgesetzte/r, KI-Verantwortliche/r, Rechtsabteilung, IT, Geschäftsleitung?

Dokumentiert? Werden wichtige Entscheidungen, Korrekturen und Blockaden dokumentiert?

MINI-ENTSCHEIDUNGSREGEL

Keine sensiblen Daten + geringe Auswirkung + Prüfung möglich → ● einfache Aufsicht ausreichend.

Kunde / HR / Finanzen / Vertrag + externe Aktion → ● menschliche Freigabe vor der Aktion zwingend.

Automatisierte Einzelentscheidung oder starke Auswirkung auf eine Person → ● doppelte Freigabe oder manuelle Bearbeitung.

Kein echtes Veto oder «dekorative» Freigabe → ● Prozess entspricht nicht dem Human-in-the-loop-Prinzip.

Zu hohes Volumen für ernsthafte Prüfung → ● Prozessdesign überarbeiten, nicht nur ein Kästchen hinzufügen.

Falsches Aufsichtssignal (Klick ohne Sicht auf die Quellen, automatischer Versand, keine Frist) → ● vor Produktivbetrieb überarbeiten.

QUELLEN
Mosqueira-Rey, E. et al. (2023). Human-in-the-loop machine learning: a state of the art. DOI: <https://doi.org/10.1007/s10462-022-10246-w>
Wu, X. et al. (2022). A survey of human-in-the-loop for machine learning. DOI: <https://doi.org/10.1016/j.future.2022.05.014>

Der Gewerbeverein — Verantwortungsvolle Praktiken
Konzipiert von Patrício André (Process AI Transformation) · CC BY-NC-SA 4.0