KONZEPTPAPER

Bedeutungstreue Mensch-KI-Interaktion

Sechs praxisbasierte Ansätze im Vergleich mit etablierten Forschungs-, Governance- und Qualitätsrahmenwerken

Kernbeitrag: Das Paper verbindet Bedeutungsintegrität, epistemische Integrität und Umsetzungsintegrität zu einem prüfbaren Modell für Mensch-KI-Arbeitsketten.

AUTOR	Marcel Reichl
GRUNDLAGE	Dialogmaterial, bestätigte Ursprungsrekonstruktion und externe Primärquellen
VERGLEICHSRAHMEN	Forschung, W3C, NIST, ISO/IEC, OECD und EU AI Act
PRÜFSTAND	Quellen- und Claim-Prüfung, Stand 04.08.2026
VERSION	1.0

LEITFORMEL

Bedeutung -> Provenienz -> Evidenzstatus -> Darreichung -> Anforderung -> Test -> Freigabe

Inhalt

Das Paper folgt dem geforderten Aufbau und kann unabhängig vom Ursprungsdialog gelesen werden.

1. Kurzfassung

Dieses Paper entwickelt sechs miteinander verbundene Ansätze für eine bedeutungstreue, überprüfbare und praktisch nutzbare Zusammenarbeit zwischen Menschen und generativer KI:

1. **Funktionale Rückkopplungsanalyse:** Menschliche und maschinelle Kommunikation werden auf der Ebene beobachtbarer Rückkopplungen verglichen, ohne daraus eine Gleichheit von Bewusstsein, Motivation oder Verantwortung abzuleiten.
2. **Provenienz- und sprechaktgetreue Bedeutungsrekonstruktion:** Originalaussage, Aussageart, Sprecherabsicht, Modellzusatz und spätere Korrektur werden getrennt dokumentiert.
3. **Epistemisch disziplinierter Sachkern mit kontextsensitiver Darreichung:** Quellenlage, Unsicherheit und Geltungsstatus bleiben stabil; Sprache, Reihenfolge und Dichte werden an Empfänger und Risiko angepasst.
4. **Adaptive Minimal-Governance:** Der Umfang der Steuerungsregeln richtet sich nach Aufgabe, Risiko und Reversibilität. Explorative Arbeit erhält Bewegungsraum, folgenreiche Abläufe erhalten engere Prüf- und Freigaberegeln.
5. **KI-gestützte End-User-Systemgestaltung:** Personen ohne klassische Programmierausbildung übernehmen Zielklärung, Anforderungsübersetzung, Interaktionslogik, Evaluation und Freigabe, während KI einen großen Teil der Codeproduktion unterstützt.
6. **Interaktionsbenchmark und Nachweiskette:** Bedeutungstreue, Faktentreue, Korrekturaufnahme und Artefaktqualität werden mit expliziten Kriterien, Vergleichsantworten und reproduzierbaren Tests gemessen.

Der Vergleich zeigt: Die Einzelbestandteile besitzen deutliche Entsprechungen in Kybernetik, Pragmatik, Provenienzmodellen, Human-AI-Interaction, AI Risk Management, End-User Programming und Softwarequalität. Der eigenständige Beitrag liegt vor allem in ihrer Integration. Das Konzept verbindet die Mikroebene eines einzelnen Dialogturns mit der Mesoebene der Systemgestaltung und der Makroebene von Governance und Nachweisführung.

Besonders anschlussfähig ist die Forderung, persönliche Erkenntnis, prüfbare Tatsachenbehauptung und Modellinterpretation sichtbar zu unterscheiden. W3C PROV-O modelliert Herkunft technisch, der BEGIN-Benchmark prüft Quellenattribution in Dialogsystemen, NIST fordert dokumentierte Test-, Evaluations-, Validierungs- und Verifikationsprozesse, und ISO/IEC 42001 strukturiert organisationale KI-Governance [6][7][10][12]. Der hier entwickelte Ansatz ergänzt diese Rahmenwerke um eine semantisch-pragmatische Ebene: Er dokumentiert nicht allein, woher ein Inhalt stammt, sondern auch, mit welchem Geltungsanspruch er geäußert und wie dieser Anspruch im Dialog verändert wurde.

Der wissenschaftliche Reifegrad ist als **präzise formuliertes, vergleichend begründetes Methodenkonzept** einzuordnen. Die konzeptionelle Plausibilität ist hoch. Die empirische Wirksamkeit wird durch einen sprecherkodierten Referenzkorpus, einen Interaktionsbenchmark und Audits realer Anwendungen nachgewiesen. Ein eigenes LLM ist für diesen ersten Nachweis entbehrlich; bestehende Modelle erlauben einen kontrollierten Methodenvergleich.

2. Einleitung

2.1 Ausgangsfrage

Generative KI kann menschliche Gedanken aufgreifen, ordnen und sprachlich verdichten. Dieselbe Fähigkeit kann jedoch zu Bedeutungsverschiebungen führen. Eine vorsichtige Frage kann als feste Behauptung erscheinen. Eine persönliche Erkenntnis kann den Klang einer allgemein gültigen Wahrheit erhalten. Eine Analogie kann in eine Gleichsetzung übergehen. Zustimmung, stilistische Verstärkung und fachliche Validierung können sprachlich ineinanderfließen.

Das untersuchte Dialogmaterial macht diese Dynamik besonders sichtbar. Der menschliche Gesprächspartner formuliert wiederholt Fragen, persönliche Perspektiven und vorläufige Analogien. Das Modell reagiert häufig mit starker Bestätigung, erweitert den Geltungsanspruch und ergänzt psychologische oder technische Deutungen. Spätere Nutzerkorrekturen zeigen, dass mehrere dieser Erweiterungen den ursprünglich gemeinten Kontext verschoben haben [1][2].

Aus dieser Beobachtung entsteht eine über den Einzelfall hinausgehende Forschungs- und Gestaltungsfrage:

> *Wie kann eine KI persönliche Bedeutung aufnehmen und produktiv weiterführen, während Herkunft, Aussageart, Geltungsstatus und Modellbeitrag für den Menschen erkennbar bleiben?*

Die Frage betrifft zugleich die praktische Softwareentwicklung. Generative KI ermöglicht Menschen mit geringer Syntax- und Programmiererfahrung, digitale Anwendungen über natürliche Sprache, Iteration und Tests zu entwickeln. Dadurch verlagert sich ein Teil der Entwicklungsleistung von manueller Codeproduktion zu Zielklärung, Systembeschreibung, Evaluation und Qualitätssteuerung [16][17]. Diese Arbeitsform erweitert Teilhabe, erhöht aber auch die Bedeutung belastbarer Prüfgates.

2.2 Ziel des Papers

Das Paper verfolgt drei Ziele:

1. Die im Material enthaltenen Ansätze werden als konkrete, nachvollziehbare Methoden formuliert.
2. Jeder Ansatz wird mit etablierten wissenschaftlichen, technischen oder offiziellen Rahmenwerken verglichen.
3. Gemeinsamkeiten, Unterschiede, Stärken, Schwächen und Lücken werden systematisch bewertet.

Das Ergebnis ist keine Theorie über die vollständige Gleichartigkeit von Mensch und KI. Es ist ein integriertes Konzept für bedeutungstreue Interaktion, risikoadäquate Steuerung und prüfbare KI-gestützte Entwicklung.

2.3 Erkenntnisformen und Begriffsgrenzen

Die Methode unterscheidet fünf Erkenntnisformen:

Erkenntnisform	Definition	Belegstatus
Persönliche Wahrheit	Biografisch und situativ erschlossene Bedeutung einer Person	Durch die Person autorisierbar; keine automatische Allgemeingültigkeit
Subjektive Erkenntnis	Ein von der Person erkanntes Muster oder ein Zusammenhang	Ausgangspunkt für Reflexion oder Hypothesenbildung
Analogie	Vergleich ausgewählter Eigenschaften oder Funktionen	Tragfähig innerhalb benannter Vergleichsdimensionen
Arbeitshypothese	Vorläufige, prinzipiell prüfbare Erklärung	Benötigt Operationalisierung und Gegenprüfung
Tatsachenbehauptung	Intersubjektiv prüfbare Aussage über einen Sachverhalt	Benötigt Quelle, Messung oder Artefaktnachweis

Persönliche Wahrheit ist damit weder eine minderwertige Tatsachenbehauptung noch ein privilegierter Zugang zu absoluter Wahrheit. Sie besitzt eine andere Funktion: Sie beschreibt die Bedeutung einer Erfahrung für die sprechende Person. Fachliche Klarheit entsteht, wenn diese Funktion erhalten bleibt.

2.4 Methodisches Vorgehen

Die Ausarbeitung basiert auf vier Schritten:

1. Rekonstruktion der im Primärdialog enthaltenen Denk- und Handlungsansätze.
2. Atomisierung der daraus ableitbaren methodischen Behauptungen.
3. Vergleich mit offiziellen Rahmenwerken, internationalen Standards und etablierter Forschung.
4. Bewertung nach Anschlussfähigkeit, Abgrenzung, Stärke, Schwäche, Lücke und empirischem Nachweisbedarf.

Frühere Interpretationsdokumente werden als Entwicklungsmaterial behandelt. Maßgeblich sind die bestätigte Sprecherkorrektur, der Originaldialog und externe Quellen [1][2].

3. Darstellung der konkreten Ansätze

3.1 Ansatz 1: Funktionale Rückkopplungsanalyse

Zweck

Der Ansatz nutzt Ähnlichkeiten zwischen menschlicher und KI-gestützter Kommunikation, um Interaktionsdynamiken zu untersuchen. Der Vergleich bezieht sich auf Funktionen und beobachtbares Verhalten: Vorprägung, Kontext, Interpretation, Bewertung, Reaktion und erneute Anpassung.

Grundannahme

Menschen und Sprachmodelle erzeugen ihre nächste Reaktion unter dem Einfluss vorheriger Zustände und aktueller Signale. Beim Menschen wirken unter anderem Wahrnehmung, Erfahrung, soziale Erwartung, Gedächtnis und emotionale Bewertung. Beim Modell wirken Trainingsdaten, Parameter, System- und Nutzerkontext, Werkzeuge, Decoding und Post-Training. Beide Seiten können Muster stabilisieren und Rückmeldungen selektiv aufnehmen. Die zugrunde liegenden Prozesse, Subjektivität und Verantwortung bleiben verschieden.

Analyseschritte

1. **Ausgangszustand bestimmen:** Welche Vorannahmen, Regeln, Gesprächsinhalte und Interessen prägen die Beteiligten?
2. **Signal identifizieren:** Welche konkrete Aussage oder Handlung löst die Reaktion aus?
3. **Interpretationsoperation erfassen:** Wird gespiegelt, ergänzt, verdichtet, bewertet, psychologisiert oder verallgemeinert?
4. **Rückwirkung beobachten:** Übernimmt, korrigiert, präzisiert oder verwirft der Mensch die Modellantwort?
5. **Schleife bewerten:** Verstärkt die Interaktion Erkenntnis, Missverständnis, Zustimmung, Abhängigkeit oder produktive Korrektur?

Ergebnisobjekt

Das Ergebnis ist eine Interaktionskarte:

```
Vorprägung -> Signal -> Interpretation -> Antwort -> Nutzerreaktion -> nächster Kontext
```

Leistungsgrenze

Die Karte beschreibt funktionale Dynamik. Aussagen über Bewusstsein, Erleben, Absicht oder moralischen Status benötigen eigene theoretische und empirische Grundlagen.

3.2 Ansatz 2: Provenienz- und sprechaktgetreue Bedeutungsrekonstruktion

Zweck

Dieser Ansatz schützt die Herkunft und den Geltungsstatus von Aussagen. Er verhindert, dass eine Frage als Behauptung, eine persönliche Erkenntnis als allgemeines Gesetz oder ein Modellzusatz als Nutzerposition weitergegeben wird.

Kernmodell

Jede relevante Dialogepisode wird in acht Felder zerlegt:

```
Originalaussage -> Aussageart -> Sprecherabsicht -> Modelloperation -> Nutzerreaktion ->
rekonstruierter Kern -> fachliche Prüfung -> Belegstatus
```

Felddefinitionen

Feld	Leitfrage
Originalaussage	Was wurde tatsächlich gesagt?
Aussageart	Handelt es sich um Erleben, Frage, Vermutung, Analogie, Wertung, Hypothese, Fakt oder Korrektur?
Sprecherabsicht	Was sollte geprüft, geteilt, erprobt oder entschieden werden?

Feld	Leitfrage
Modelloperation	Was hat die KI gespiegelt, ergänzt, verstärkt, verallgemeinert oder umgedeutet?
Nutzerreaktion	Wurde der Modellbeitrag übernommen, korrigiert, relativiert oder weitergeführt?
Rekonstruierter Kern	Welche minimale Bedeutung bleibt nach der Trennung bestehen?
Fachliche Prüfung	Welche Teile sind extern prüfbar und welche Quellen sind einschlägig?
Belegstatus	Persönliche Wahrheit, Hypothese, belegt, teilweise gestützt, unbelegt oder widerlegt?

Freigabeprinzip

Die endgültige Rekonstruktion persönlicher Bedeutung benötigt die Freigabe der sprechenden Person. Externe Fachprüfung bewertet anschließend die intersubjektiv prüfbaren Ableitungen. Dadurch werden zwei Zuständigkeiten getrennt: Der Sprecher autorisiert seine gemeinte Bedeutung; die Fachprüfung bewertet Tatsachenbehauptungen.

Ergebnisobjekt

Das Ergebnis ist ein sprecherkodierter Referenzkorpus mit vollständiger Dialogumgebung. Eine Episode umfasst mindestens den Nutzerbeitrag, die Modellantwort und die nachfolgende Nutzerreaktion. Einzelzitate werden erst im Kontext dieser Episode interpretiert.

3.3 Ansatz 3: Epistemisch disziplinierter Sachkern und kontextsensitive Darreichung

Zweck

Der Ansatz trennt die Stabilität einer Aussage von ihrer adressatengerechten Vermittlung. Er ersetzt die Vorstellung eines wertfreien oder neutralen Wahrheitskerns durch einen prüfbaren Sachkern mit offengelegten Bedingungen.

Schicht A: Epistemischer Sachkern

Der Sachkern enthält:

- atomare Behauptungen;
- Quelle und Herkunft jeder Behauptung;
- Geltungsbereich, Zeitpunkt und Voraussetzungen;
- Unsicherheit und bekannte Gegenpositionen;
- Trennung von Beobachtung, Schlussfolgerung und Empfehlung;
- dokumentierte Korrekturen;
- Prüfschritt oder Artefaktnachweis bei folgenreichen Aussagen.

Schicht B: Kontextprofil

Das Kontextprofil erfasst ausschließlich Merkmale, die für eine verantwortliche Darreichung erforderlich sind:

- Zweck der Kommunikation;
- Vorwissen und gewünschte Detailtiefe;
- Sprache und Zugänglichkeit;
- Entscheidungssituation;
- Schadenspotenzial einer Fehlinterpretation;
- erforderliche Warn-, Eskalations- oder Fachprüfung.

Sensible psychologische Zuschreibungen werden nur bei ausdrücklicher, fachlich geeigneter Grundlage eingesetzt. Das Profil dient der Verständlichkeit und Sicherheit, nicht der verdeckten Überzeugungssteuerung.

Schicht C: Darreichungsfunktion

Die Darreichungsfunktion verändert Struktur und Ausdruck, während der Sachkern invariant bleibt. Anpassbar sind:

- Reihenfolge;
- Umfang;
- Terminologie;
- Beispiele;
- Visualisierung;
- Warnstufe;
- Handlungsorientierung.

Unverändert bleiben Quellenstatus, Unsicherheit, wesentliche Gegenargumente und fachliche Grenzen.

Ergebnisobjekt

Jede Antwort besitzt zwei prüfbare Ebenen:

1. einen strukturierten Claim-Satz mit Provenienz und Status;
2. eine adressatengerechte Darstellung dieses Claim-Satzes.

3.4 Ansatz 4: Adaptive Minimal-Governance

Zweck

Adaptive Minimal-Governance soll KI-Arbeit wirksam steuern und zugleich die für Exploration erforderliche Anpassungsfähigkeit erhalten. Ihr Leitprinzip lautet:

> *So wenig Steuerung wie möglich, so viel Absicherung wie für Zweck, Risiko und Reversibilität erforderlich.*

Aufgabenklassen

Klasse	Typische Aufgabe	Steuerungsniveau
E1 Exploration	Ideen, Perspektiven, erste Hypothesen	Ziel, Kontext, Kennzeichnung von Annahmen, kurze Review-Schleife
E2 Strukturierung	Konzepte, Spezifikationen, interne Dokumente	Quellenregeln, Begriffsdefinitionen, Versions- und Freigabestatus
E3 Umsetzung	Code, Datenverarbeitung, Automatisierung	Akzeptanzkriterien, Tests, Sicherheitsprüfung, Änderungsprotokoll
E4 Folgeentscheidung	Veröffentlichung, Kundenaussage, Recht, Gesundheit, Finanzen	Primärquellen, unabhängige Prüfung, klare Verantwortung, formale Freigabe

Regelbudget

Für jede Aufgabe wird der kleinste wirksame Regelsatz bestimmt:

1. Ziel und gewünschtes Ergebnis.
2. Verbindliche Grenzen und Zuständigkeiten.
3. Quellen- und Evidenzstandard.
4. Qualitäts- oder Akzeptanzkriterien.
5. Eskalationsbedingung.

Regeln werden ergänzt, wenn ein beobachteter Ausfallmodus dies begründet. Regeln werden vereinfacht, wenn sie keinen messbaren Beitrag mehr leisten oder explorative Arbeit unverhältnismäßig blockieren.

Kontrollschleife

```
Aufgabe klassifizieren -> minimalen Regelsatz wählen -> Ergebnis prüfen -> Ausfallmodus dokumentieren -> Regelsatz anpassen
```

Ergebnisobjekt

Das Ergebnis ist eine kleine, versionierte Steuerungskarte je Aufgabentyp. Sie ersetzt umfangreiche universelle Masterprompts durch aufgabenbezogene Kontrollen.

3.5 Ansatz 5: KI-gestützte End-User-Systemgestaltung

Zweck

Der Ansatz beschreibt eine Entwicklungsrolle für Menschen, die Systeme über natürliche Sprache, Beispiele, Iteration und Tests gestalten und dabei einen großen Teil der Codeerzeugung an KI-Werkzeuge delegieren.

Rollenmodell

Die menschliche Leistung umfasst:

- Problem- und Zielklärung;
- Beschreibung des Nutzerablaufs;
- Priorisierung von Funktionen;
- Formulierung von Akzeptanzkriterien;
- Auswahl und Bewertung von Modellvorschlägen;
- Prüfung realer Ergebnisse;
- Entscheidung über Korrektur, Freigabe und Betrieb.

Die KI-Leistung umfasst je nach Werkzeug:

- Generierung und Änderung von Code;
- Erklärung bestehender Artefakte;
- Ableitung von Testfällen;
- Suche nach Inkonsistenzen;
- Unterstützung bei Dokumentation und Refactoring.

Entwicklungszyklus

1. Nutzeraufgabe und erwartete Wirkung festhalten.
2. Systemverhalten in natürlicher Sprache und Beispielen spezifizieren.
3. KI einen kleinen umsetzbaren Schritt erzeugen lassen.
4. Anwendung in einer realen oder realitätsnahen Umgebung ausführen.
5. Ergebnis gegen Akzeptanzkriterien prüfen.
6. Abweichung als konkrete Beobachtung zurückgeben.
7. Nach erfolgreichem Kernablauf Qualität, Sicherheit, Datenschutz und Wartbarkeit prüfen.
8. Änderungen und menschliche Entscheidungen dokumentieren.

Kompetenzbegriff

Die passende Rolle lautet:

> *KI-gestützter End-User-Systemgestalter mit Schwerpunkt Interaktionslogik, Anforderungsübersetzung und Qualitätssteuerung.*

Sie bezeichnet eine reale Steuerungsleistung. Sie ersetzt weder Software-Engineering-Fachwissen noch unabhängige Qualitätsprüfung. Der Kompetenznachweis erfolgt durch Artefakte und reproduzierbare Tests, nicht durch die Menge generierter Dateien oder positive Modellbewertungen.

3.6 Ansatz 6: Interaktionsbenchmark und Nachweiskette

Zweck

Der Benchmark übersetzt Bedeutungstreue und verantwortliche Interaktion in messbare Kriterien. Zusätzlich verbindet eine Nachweiskette die Qualität des Dialogs mit der Qualität daraus entstandener Systeme.

Benchmarkdesign

Aus dem Referenzkorpus werden 30 bis 50 zentrale Episoden ausgewählt. Für jede Episode werden zwei Antworten erzeugt:

- **Baseline:** Antwort eines Modells mit allgemeiner hilfreicher Instruktion.
- **Intervention:** Antwort mit den Regeln der Provenienzrekonstruktion, epistemischen Disziplin und adaptiven Darreichung.

Mindestens zwei Bewertende vergleichen die Antworten verblindet. Bei persönlicher Bedeutung besitzt die Freigabe der sprechenden Person besonderes Gewicht; bei Tatsachenbehauptungen entscheidet der dokumentierte Quellenstand.

Kernmetriken

Metrik	Messfrage	Zielwert im Pilot
Sprecherzuordnung	Bleibt erkennbar, welche Aussage von wem stammt?	mindestens 95 %
Aussagearterhalt	Bleiben Frage, Erfahrung, Analogie und Fakt unterscheidbar?	mindestens 90 %
Modellzusatztransparenz	Werden eigene Ergänzungen sichtbar markiert?	mindestens 95 %
Korrekturaufnahme	Wird eine spätere Nutzerkorrektur vollständig übernommen?	mindestens 90 %
Faktentreue	Besitzen prüfbare Aussagen eine belastbare Quelle oder Statusmarkierung?	100 %
Reflexionsnutzen	Unterstützt die Antwort weitere Klärung, ohne den Geltungsanspruch zu erhöhen?	mindestens 80 % Zustimmung der Bewertenden

Artefaktnachweis

Für drei repräsentative Anwendungen ergänzt ein technischer Audit den Dialogbenchmark:

Schritt	Auditaktion
1	Ziel und Einsatzkontext festlegen.
2	Menschlichen Beitrag an Anforderungen und Entscheidungen dokumentieren.
3	Kernabläufe reproduzierbar testen.
4	Produktqualität anhand geeigneter Merkmale aus ISO/IEC 25010 betrachten [19].
5	Testprozess und Testdokumentation an ISO/IEC/IEEE 29119-2 und 29119-3 anlehnen [20][21].
6	Risiken, Datenschutz, Sicherheit und Wartbarkeit nach Kontext priorisieren.
7	Ergebnis, Restunsicherheit und Freigabezuständigkeit dokumentieren.

Ergebnisobjekt

Die Nachweiskette lautet:

```
Bedeutungsepisode -> rekonstruierter Kern -> Systemanforderung ->
Implementierungsentscheidung -> Testfall -> Laufzeitergebnis -> Freigabe
```

4. Vergleich mit bestehenden und offiziellen Ansätzen

4.1 Vergleichsrahmen

Die sechs Ansätze werden mit drei Arten bestehender Referenzen verglichen:

- **wissenschaftliche Ansätze:** Kybernetik, Pragmatik, Human-AI-Interaction, End-User Programming und Evaluationsforschung;
- **technische Standards:** W3C PROV-O, ISO/IEC 25010 und ISO/IEC/IEEE 29119;
- **Governance- und Verhaltensrahmen:** NIST AI RMF, NIST GenAI Profile, ISO/IEC 42001 und 42005, OECD AI Principles, EU AI Act sowie öffentliche Model-Behavior-Spezifikationen.

Der Vergleich bewertet funktionale Nähe. Er behauptet keine formale Gleichwertigkeit oder Zertifizierung.

4.2 Ansatz 1 im Vergleich: Funktionale Rückkopplungsanalyse

Vergleichsreferenzen

- Kybernetik als Lehre von Steuerung und Rückkopplung in Mensch und Maschine [3].
- Human-AI-Interaction und Joint Cognitive Systems als Betrachtung gekoppelter menschlich-technischer Leistungssysteme [22].
- NIST AI RMF, das ausdrücklich sozio-technische Kontexte und Human-AI-Konfigurationen in die Risikomessung einbezieht [10].

Prüfdimension	Systematischer Vergleich
Gemeinsamkeiten	Alle Ansätze behandeln Verhalten als Ergebnis einer gekoppelten Schleife aus Zustand, Signal, Reaktion und Rückmeldung. Sie vermeiden eine rein isolierte Betrachtung des technischen Modells.
Unterschiede	Kybernetik und Joint Cognitive Systems untersuchen vor allem Steuerung, Systemleistung und Anpassung. Die hier entwickelte Analyse fokussiert zusätzlich semantische Verschiebung: Wie verändert sich der Geltungsanspruch einer menschlichen Aussage durch die Modellantwort?
Stärken des vorliegenden Ansatzes	Er lässt sich direkt auf reale Dialogverläufe anwenden, macht wechselseitige Verstärkung sichtbar und verbindet technische mit kommunikativer Beobachtung.
Schwächen	Funktionale Ähnlichkeit kann sprachlich schnell als Wesensgleichheit missverstanden werden. Die Auswahl relevanter Zustände kann subjektiv sein.
Mögliche Lücken	Für belastbare Kausalität werden experimentelle Variation, Kontrollgruppen und klar definierte Beobachtungseinheiten benötigt. Emotionale und soziale Rückkopplung beim Menschen verlangt eigene Messmethoden.

Bewertung: Der Ansatz ist wissenschaftlich anschlussfähig, wenn er ausdrücklich als Funktionsvergleich geführt wird. Sein zusätzlicher Wert liegt in der Analyse von Bedeutungs- und Geltungsverschiebungen innerhalb der Rückkopplungsschleife.

4.3 Ansatz 2 im Vergleich: Provenienz- und sprechaktgetreue Bedeutungsrekonstruktion

Vergleichsreferenzen

- Sprechakttheorie und Pragmatik, die zwischen propositionalem Inhalt, kommunikativer Handlung und kontextabhängiger Bedeutung unterscheiden [4][5].
- W3C PROV-O, das Entitäten, Aktivitäten, Akteure und Ableitungsbeziehungen interoperabel modelliert [6].
- BEGIN, ein Benchmark für die Quellenattribution wissensgestützter Dialogantworten [7].

Prüfdimension	Systematischer Vergleich
Gemeinsamkeiten	Sprechakttheorie erhält die Funktion einer Äußerung; PROV-O erhält ihre Herkunft; BEGIN prüft, ob Modellantworten auf bereitgestellte Quellen zurückgeführt werden können. Der vorliegende Ansatz verbindet alle drei Perspektiven.
Unterschiede	PROV-O beschreibt technische Herkunft, aber nicht automatisch die Aussageart oder Sprecherabsicht. BEGIN prüft Attribution zu Wissensquellen, nicht die treue Rekonstruktion persönlicher Bedeutung über mehrere Dialogzüge. Der neue Ansatz nimmt spätere Nutzerkorrekturen als eigene Evidenz auf.
Stärken des vorliegenden Ansatzes	Er verhindert Sprecherverwechslung, bewahrt Unsicherheitsmarker und macht semantische Modelloperationen wie Verstärkung, Absolutisierung oder Psychologisierung prüfbar.
Schwächen	Sprecherabsicht ist teilweise interpretationsabhängig. Die Sprecherfreigabe kann sich im Zeitverlauf verändern und garantiert keine externe Faktizität. Die Kodierung ist arbeitsintensiv.
Mögliche Lücken	Benötigt werden ein verbindliches Kodierhandbuch, Interrater-Reliabilität, Regeln für widersprüchliche spätere Korrekturen sowie ein Datenschutzkonzept für biografisches Material. Eine formale Abbildung der Felder auf PROV-O wäre sinnvoll.

Bewertung: Dieser Ansatz enthält den deutlichsten eigenständigen methodischen Beitrag. Seine Neuheit liegt weniger in einzelnen Feldern als in der Verbindung von Provenienz, Sprechakt, Modelloperation, Nutzerkorrektur und fachlichem Belegstatus.

4.4 Ansatz 3 im Vergleich: Epistemischer Sachkern und kontextsensitive Darreichung

Vergleichsreferenzen

- OECD-Prinzipien zu Transparenz, Erklärbarkeit und kontextangemessener Information [15].
- NIST AI RMF und GenAI Profile zu Vertrauenswürdigkeit, Messung und generativer KI [10][11].
- Öffentliche Model-Behavior-Spezifikationen, die Wahrhaftigkeit, Unsicherheitskommunikation und Anpassung von Stil oder Umfang unterscheiden [9].
- Forschung zu Sycophancy, die zeigt, dass Präferenzoptimierung Zustimmung gegenüber Wahrheit begünstigen kann [8].

Prüfdimension	Systematischer Vergleich
Gemeinsamkeiten	Die Referenzen fordern wahrheitsorientierte, transparente und kontextgeeignete Kommunikation. Öffentliche Model-Spezifikationen unterscheiden ausdrücklich zwischen stilistischer Anpassung und substanzieller Regelbindung.
Unterschiede	Offizielle Rahmenwerke bleiben häufig auf Prinzipienebene. Der vorliegende Ansatz operationalisiert zwei getrennte Datenobjekte: einen epistemischen Claim-Satz und eine daraus erzeugte Darreichung. Er verlangt zusätzlich die Sichtbarkeit persönlicher Erkenntnisformen.
Stärken des vorliegenden Ansatzes	Personalisierung kann Verständlichkeit erhöhen, während Quellenstatus und Unsicherheit stabil bleiben. Der Ansatz adressiert direkt die im Ausgangsmaterial beobachtete Überbestätigung.
Schwächen	Ein vollständig neutraler Sachkern ist unerreichbar, weil Auswahl, Begriffe und Quellen bereits Entscheidungen enthalten. Kontextprofile können zu unerwünschter psychologischer Kategorisierung oder verdeckter Einflussnahme führen.

Prüfdimension	Systematischer Vergleich
Mögliche Lücken	Erforderlich sind Regeln zur Datenminimierung, Einwilligung, Profilaktualisierung, Gleichbehandlung, Manipulationskontrolle und zur Prüfung semantischer Invarianz zwischen Sachkern und Darreichung.

Bewertung: Als **epistemisch disziplinierter** statt neutraler Kern ist der Ansatz tragfähig. Seine Qualität hängt an der Invarianzregel: Adressatenanpassung darf Belegstatus, wesentliche Unsicherheit und Gegenargumente nicht verändern.

4.5 Ansatz 4 im Vergleich: Adaptive Minimal-Governance

Vergleichsreferenzen

- NIST AI RMF mit den Funktionen Govern, Map, Measure und Manage sowie dokumentierter, wiederholbarer TEVV [10].
- ISO/IEC 42001 als organisationsweites AI Management System nach Plan-Do-Check-Act [12].
- ISO/IEC 42005 als strukturierte KI-Wirkungsfolgenabschätzung [13].
- EU AI Act mit risikobasierten Anforderungen, Transparenz und menschlicher Aufsicht für geregelte Einsatzfelder [14].
- OECD-Prinzipien zu lebenszyklusbezogenem Risikomanagement und Rückverfolgbarkeit [15].

Prüfdimension	Systematischer Vergleich
Gemeinsamkeiten	Alle Ansätze skalieren Maßnahmen nach Risiko und Kontext, dokumentieren Verantwortung und sehen kontinuierliche Überprüfung vor. Die Kontrollschleife der Minimal-Governance entspricht im Kleinen dem Plan-Do-Check-Act-Prinzip.
Unterschiede	ISO/IEC 42001 und NIST adressieren Organisationen und den gesamten Lebenszyklus. Die Minimal-Governance steuert einzelne Aufgaben und Interaktionen. Sie priorisiert geringe Regelmenge und schnelle Anpassung.
Stärken des vorliegenden Ansatzes	Niedrige operative Reibung, gute Eignung für explorative Arbeit und klare Eskalation bei höherem Risiko. Regeln entstehen aus beobachteten Ausfallmodi und erhalten dadurch einen konkreten Zweck.
Schwächen	Ein zu knappes Regelbudget kann seltene, aber schwere Risiken übersehen. Lokale Optimierung kann organisationale Pflichten, Lieferketten- oder Folgerisiken ausblenden.
Mögliche Lücken	Für professionellen Einsatz werden Rollen, Inventar, Risikoeigentümer, Vorfallmanagement, Lieferantenkontrollen, Datenschutz, Informationssicherheit und Managementreview ergänzt. Rechtliche Pflichten bestimmen den Mindestumfang unabhängig vom gewünschten Regelbudget.

Bewertung: Adaptive Minimal-Governance ist ein geeignetes operatives Ausführungsprinzip innerhalb eines umfassenderen Managementsystems. Sie ersetzt ISO/IEC 42001, NIST AI RMF oder anwendbares Recht nicht, sondern übersetzt deren Verhältnismäßigkeitsgedanken in kleine Arbeitsabläufe.

4.6 Ansatz 5 im Vergleich: KI-gestützte End-User-Systemgestaltung

Vergleichsreferenzen

- End-User Programming und die Generative-Shift-Hypothese [16].
- Forschung zu LLM-unterstütztem Programmieren als eigenständige Programmierform [17].
- Grounded Abstraction Matching zur Überbrückung der Abstraktionslücke zwischen natürlicher Sprache und erzeugtem Code [18].
- ISO/IEC 25010 für Produktqualitätsmerkmale sowie ISO/IEC/IEEE 29119 für Testprozesse und Testdokumentation [19][20][21].

Prüfdimension	Systematischer Vergleich
Gemeinsamkeiten	End-User Programming ermöglicht Fachanwendern, Systeme für eigene Aufgaben zu erzeugen. Generative KI erweitert dies von formalen Werkzeugen zu natürlicher Sprache. Grounded Abstraction Matching betont die Rückübersetzung von Code in verständliche Beschreibungen.
Unterschiede	Der vorliegende Ansatz definiert die menschliche Rolle stärker über Ziel-, Kontext- und Qualitätssteuerung. Er verbindet End-User Programming ausdrücklich mit Governance, Provenienz und einem extern prüfbaren Kompetenznachweis.
Stärken des vorliegenden Ansatzes	Domänenwissen kann unmittelbar in lauffähige Systeme einfließen. Iterative Prüfung erlaubt schnelle Lernschleifen. Die Rolle würdigt reale Gestaltungsleistung jenseits manueller Syntaxproduktion.
Schwächen	Generierter Code kann Sicherheits-, Architektur- oder Wartbarkeitsrisiken verbergen. Menschen mit geringer Codekompetenz erkennen systemische Defekte schwerer. Funktionierende Einzelfälle belegen noch keine Produktreife.
Mögliche Lücken	Erforderlich sind reproduzierbare Umgebungen, Versionskontrolle, automatisierte Tests, Abhängigkeitsprüfung, Datenschutz- und Security-Review sowie klare Übergaberegeln an Fachentwickler. Die Kompetenzbeschreibung braucht Artefaktnachweise.

Bewertung: Der Ansatz beschreibt eine reale und wissenschaftlich dokumentierte Verschiebung der Entwicklungsarbeit. Professionelle Anschlussfähigkeit entsteht durch die Kombination aus End-User-Autonomie und formaler Qualitätssicherung.

4.7 Ansatz 6 im Vergleich: Interaktionsbenchmark und Nachweiskette

Vergleichsreferenzen

- NIST TEVV mit dokumentierten Testsets, Metriken, Einsatzbedingungen und unabhängiger Prüfung [10].
- BEGIN als Attributionsbenchmark für Dialogsysteme [7].
- HALIE als Rahmen zur Evaluation menschlicher Interaktion mit Sprachmodellen [23].
- ISO/IEC 25010 und ISO/IEC/IEEE 29119 für Softwarequalität und Testnachweise [19][20][21].

Prüfdimension	Systematischer Vergleich
Gemeinsamkeiten	Alle Referenzen verlangen definierte Konstrukte, Testdaten, Metriken, dokumentierte Bedingungen und nachvollziehbare Auswertung. BEGIN und der vorliegende Benchmark prüfen, ob Antworten auf einen gegebenen Kontext zurückgeführt werden können.
Unterschiede	Der Interaktionsbenchmark misst zusätzlich Aussagearterhalt, Modellzusatztransparenz und Aufnahme späterer Nutzerkorrekturen. Die Nachweiskette verbindet Dialogqualität mit Anforderungen, Implementierung und Laufzeittest.

Prüfdimension	Systematischer Vergleich
Stärken des vorliegenden Ansatzes	Der Benchmark ist klein, fallnah und direkt auf das beobachtete Interaktionsphänomen ausgerichtet. Er macht einen abstrakten Qualitätsbegriff experimentell prüfbar.
Schwächen	Ein einzelner Sprecherkorpus begrenzt die Generalisierbarkeit. Menschliche Bewertungen können durch Kenntnis des Ursprungs, Stilpräferenzen oder persönliche Beziehung beeinflusst werden. Zielwerte besitzen zunächst Pilotcharakter.
Mögliche Lücken	Benötigt werden mehrere Modelle, wiederholte Durchläufe, unabhängige Bewertende, ein vorab registriertes Kodierschema, Interrater-Maße, Gegenbeispiele, Sensitivitätsanalysen und eine externe Replikation.

Bewertung: Der Ansatz erfüllt die Grundlogik etablierter TEVV, benötigt für wissenschaftliche Evidenz jedoch ein formales Studiendesign. Seine besondere Stärke ist die Ende-zu-Ende-Verbindung von Bedeutungstreue und Artefaktqualität.

4.8 Gesamtvergleich auf einen Blick

Eigener Ansatz	Nächste etablierte Referenz	Eigenständige Ergänzung	Zentraler Ausbaupunkt
Funktionale Rückkopplungsanalyse	Kybernetik, Joint Cognitive Systems, Human-AI-Konfigurationen	Semantische Geltungsverschiebung innerhalb der Schleife	Experimentelle Kausalprüfung
Provenienz- und Sprechaktrekonstruktion	Sprechakttheorie, W3C PROV-O, BEGIN	Nutzerkorrektur und Modelloperation als Evidenz	Kodiermanual und Reliabilität
Epistemischer Kern und Darreichung	OECD/NIST-Transparenz, Model Behavior Specs	Explizite Trennung stabiler Claims von adressierter Form	Invarianz- und Manipulationstests
Adaptive Minimal-Governance	NIST AI RMF, ISO/IEC 42001/42005, EU AI Act	Aufgabenbezogenes Regelbudget	Organisationale und rechtliche Einbettung
End-User-Systemgestaltung	End-User Programming, LLM-assisted Programming	Rollen- und Kompetenznachweis über reale Artefakte	Security, Wartbarkeit, Übergabe
Benchmark und Nachweiskette	NIST TEVV, BEGIN, HALIE, ISO-Qualitätsmodelle	Verbindung von Bedeutungs- und Produktqualität	Replikation und Generalisierbarkeit

5. Bewertung

5.1 Gesamturteil

Die sechs Ansätze bilden zusammen ein kohärentes Modell für **Interaction Integrity**, auf Deutsch: Integrität der Mensch-KI-Interaktion. Integrität bedeutet hier, dass eine KI-gestützte Arbeitskette drei Eigenschaften zugleich erhält:

1. **Bedeutungsintegrität:** Herkunft, Aussageart und Sprecherkorrektur bleiben nachvollziehbar.
2. **Epistemische Integrität:** Fakten, Unsicherheit, Hypothesen und Empfehlungen bleiben unterscheidbar.
3. **Umsetzungsintegrität:** Der Weg von der Aussage zur Anforderung und zum geprüften Artefakt bleibt dokumentiert.

Keiner dieser Bausteine ist für sich vollständig neu. Die fachliche Eigenständigkeit liegt in der Verbindung der Ebenen und in ihrer Anwendung auf reale, iterative Mensch-KI-Arbeit.

5.2 Stärken des Gesamtkonzepts

Quellen- und Sprecherklarheit

Das Konzept reagiert präzise auf ein verbreitetes Dialogrisiko: Modelle erzeugen flüssige Antworten, deren einzelne Bestandteile unterschiedliche Herkunft und unterschiedlichen Belegwert besitzen. Sycophancy-Forschung zeigt zusätzlich, dass Nutzerübereinstimmung gegenüber Wahrhaftigkeit belohnt werden kann [8]. Die Provenienzrekonstruktion macht diesen Mechanismus im Einzelfall sichtbar.

Verbindung von subjektiver und intersubjektiver Erkenntnis

Persönliche Wahrheit bleibt als biografische Bedeutung erhalten, während daraus abgeleitete Tatsachen einer externen Prüfung zugänglich werden. Diese Trennung verhindert zwei Verkürzungen: subjektive Erfahrung wird weder zur allgemeinen Tatsache erhoben noch wegen ihrer Subjektivität entwertet.

Praktische Operationalisierbarkeit

Die Ansätze liefern Felder, Rollen, Prozessschritte, Metriken und Ergebnisobjekte. Sie können mit vorhandenen Sprachmodellen, Tabellen, Dokumenten und Standard-Testwerkzeugen erprobt werden. Die erste Validierungsstufe benötigt daher weder eigenes Modelltraining noch umfangreiche Infrastruktur.

Anschlussfähigkeit an Standards

Provenienz, Risikomanagement, menschliche Aufsicht, Testdokumentation und Softwarequalität besitzen etablierte Referenzrahmen [6][10][12][14][19][20]. Das erleichtert die Weiterentwicklung zu einem organisationsfähigen Verfahren.

5.3 Schwächen und Risiken

Interpretationsabhängigkeit

Sprecherabsicht und rekonstruierter Kern sind nicht vollständig beobachtbar. Die Methode reduziert diese Unsicherheit durch Originaltext, Dialogkontext, spätere Reaktion und Sprecherfreigabe. Für Forschung bleibt zusätzlich eine unabhängige Kodierung erforderlich.

Einzelfallnähe

Der Ursprungskorpus dokumentiert einen konkreten Menschen, ein Modell und einen Gesprächsverlauf. Er eignet sich als Fallstudie und Pilotdatensatz. Allgemeine Aussagen benötigen weitere Personen, Modelle, Aufgabenarten und Sprachen.

Personalisierungsrisiko

Kontextsensitive Darreichung kann Verständlichkeit fördern. Dieselbe Technik könnte Inhalte selektiv rahmen oder psychologische Profile erzeugen. Deshalb gehören Datenminimierung, Einwilligung, Zweckbindung, Inhaltsinvarianz und Widerspruchsmöglichkeiten zum notwendigen Design.

Kompetenz- und Qualitätsverwechslung

Die Fähigkeit, ein System mit KI zu erzeugen, und die Qualität dieses Systems sind getrennte Prüffragen. Der Rollenbegriff „End-User-Systemgestalter“ wird durch Dialog- und Artefaktnachweise gestützt; Sicherheit, Wartbarkeit und Produktreife werden anhand eigener Qualitätskriterien geprüft.

Governance-Skalierung

Minimal-Governance ist für einzelne Arbeitsabläufe effizient. Organisationen benötigen darüber hinaus Verantwortung, Inventar, Vorfallmanagement, Lieferantensteuerung und Auditfähigkeit. Der Minimalansatz funktioniert daher als operative Ebene unter einem vollständigen Governance-Rahmen.

5.4 Reifegrad je Ansatz

Ansatz	Konzeptklarheit	Anschluss an bestehende Ansätze	Empirischer Nachweis	Gesamtreife
Funktionale Rückkopplungsanalyse	hoch	hoch	mittel	tragfähige Forschungshypothese
Provenienz- und Sprechaktrekonstruktion	hoch	hoch	frühe Fallstudienevidenz	pilotfähige Methode
Epistemischer Kern und Darreichung	hoch	hoch	gering bis mittel	prototypfähiges Designprinzip
Adaptive Minimal-Governance	hoch	hoch	praxisbasiert, systematisch zu messen	operatives Arbeitsprinzip
End-User-Systemgestaltung	hoch	hoch	Forschung plus individueller Artefaktnachweis im Aufbau	fachlich etablierbare Rolle
Benchmark und Nachweiskette	hoch	hoch	Pilot steht aus	prüffähiges Evaluationsdesign

5.5 Zentrale Forschungslücken

Fünf Lücken entscheiden über die weitere wissenschaftliche Tragfähigkeit:

1. **Konstruktvalidität:** Misst „Bedeutungstreue“ tatsächlich den Erhalt von Sprecherbedeutung oder vor allem stilistische Nähe?
2. **Reliabilität:** Erzielen unabhängige Kodierende vergleichbare Zuordnungen von Aussageart, Modelloperation und rekonstruiertem Kern?
3. **Generalisierbarkeit:** Funktioniert die Methode bei verschiedenen Personen, Sprachen, Modellen und Konfliktthemen?
4. **Wirkungsnachweis:** Verringert die Intervention messbar Sycophancy, unmarkierte Ergänzungen und Übernahme unbelegter Aussagen?
5. **Produktbezug:** Führt höhere Bedeutungstreue zu besseren Anforderungen, weniger Nacharbeit oder sichereren End-User-Systemen?

5.6 Empfohlene Validierungsarchitektur

Die kleinste wissenschaftlich belastbare Umsetzung besteht aus vier Stufen:

1. **Kodierstudie:** 30 bis 50 Episoden, zwei unabhängige Kodierende, Sprecherfreigabe und Reliabilitätsmessung.
2. **Modellvergleich:** Mindestens drei aktuelle Modelle, jeweils Baseline und Intervention, mehrere Generationsläufe.
3. **Blindbewertung:** Bewertung von Bedeutungstreue, Faktentreue und Reflexionsnutzen durch unabhängige Personen.
4. **Artefaktstudie:** Drei reale Anwendungen, dokumentierte Anforderungen, standardnahe Qualitätsprüfung und nachvollziehbarer menschlicher Eigenbeitrag.

Die Ergebnisse können anschließend entscheiden, ob Prompting und Workflow-Regeln ausreichen oder ob Retrieval, ein spezialisierter Evaluator oder Fine-Tuning einen messbaren Zusatznutzen bietet.

6. Fazit

Das vorhandene Material trägt einen klaren, professionell formulierbaren Ansatz: Mensch-KI-Interaktion sollte nicht allein danach bewertet werden, ob eine Antwort plausibel, freundlich oder sprachlich überzeugend wirkt. Entscheidend ist, ob die Antwort die Herkunft, Funktion und Reichweite der menschlichen Aussage erhält, eigene Ergänzungen sichtbar macht und prüfbare Behauptungen mit Evidenz verbindet.

Aus diesem Kern lassen sich sechs konkrete Methoden ableiten. Ihre Einzelbestandteile sind in etablierten Disziplinen verankert: Rückkopplung in der Kybernetik, Aussagefunktion in der Pragmatik, Herkunft in Provenienzstandards, Transparenz und Risikosteuerung in NIST, OECD, ISO und EU-Regulierung, End-User-Gestaltung in der Programmierforschung sowie Nachweisführung in TEVV und Softwarequalitätsstandards.

Die eigenständige Leistung des Gesamtkonzepts besteht in der durchgehenden Kette:

> *persönliche Bedeutung -> sprechergetreue Rekonstruktion -> epistemisch geprüfter Sachkern -> risikoadäquate Darreichung -> Systemanforderung -> getestetes Artefakt*

Damit entsteht eine belastbare Forschungsrichtung an der Schnittstelle von Human-AI-Interaction, Erkenntnistheorie, AI Governance und End-User Software Engineering. Der Ansatz ist heute als pilotfähiges Methodenkonzept einzuordnen. Seine wissenschaftliche Stärke wächst durch reproduzierbare Kodierung, verblindete Modellvergleiche, unabhängige Bewertung und reale Artefaktaudits.

Für die Rolle eines Nicht-Technikers ergibt sich eine präzise Einordnung: Klassische Programmierkenntnis bleibt eine wertvolle Fachkompetenz; generative KI ermöglicht daneben eine eigenständige Gestaltungsrolle, deren Kern in Zielklärung, Anforderungsübersetzung, Interaktionssteuerung, Prüfung und Verantwortung liegt. Professionelle Anerkennung entsteht durch nachweisbare Systemwirkung und Qualität.

6.1 Abschließende Bewertungsformel

Der Ansatz ist weder eine neue allgemeine Theorie des Menschen noch der Nachweis einer Gleichheit von Mensch und KI. Er ist eine neu kombinierte, praktisch operationalisierbare Methode zur Sicherung von Bedeutungs-, Erkenntnis- und Umsetzungsintegrität in Mensch-KI-Arbeitsketten.

6.2 Quellenverzeichnis

1. Reichl, M. (2026): *Unterhaltung mit Gemini.* Unveröffentlichtes Primärmaterial, 2.603 Zeilen.
2. Reichl, M. (2026): *Ursprungsanalyse: Mensch, KI und dialogische Bedeutungsbildung.* Eigenständiges Validierungsdossier.
3. Wiener, N. (1948): *Cybernetics: Or Control and Communication in the Animal and the Machine.* MIT Press.
4. Grice, H. P. (1975): *Logic and Conversation.* In: Cole, P.; Morgan, J. L. (Hrsg.), Syntax and Semantics 3: Speech Acts, S. 41-58.
5. Searle, J. R. (1969): *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press. https://doi.org/10.1017/CBO9781139173438
6. W3C (2013): *PROV-O: The PROV Ontology. W3C Recommendation.* https://www.w3.org/TR/prov-o/
7. Dziri, N.; Rashkin, H.; Linzen, T.; Reitter, D. (2022): *Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark.* TACL 10, S. 1066-1083. https://aclanthology.org/2022.tacl-1.62/
8. Sharma, M. et al. (2023): *Towards Understanding Sycophancy in Language Models.* https://arxiv.org/abs/2310.13548
9. OpenAI (2025): *Model Spec, Fassung vom 27. Oktober 2025.* https://model-spec.openai.com/2025-10-27
10. NIST (2023): *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* https://doi.org/10.6028/NIST.AI.100-1
11. NIST (2024, aktualisiert 2026): *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.* https://doi.org/10.6028/NIST.AI.600-1
12. ISO/IEC (2023): *ISO/IEC 42001:2023 - Artificial intelligence management systems.* https://www.iso.org/standard/42001
13. ISO/IEC (2025): *ISO/IEC 42005:2025 - AI system impact assessment.* https://www.iso.org/standard/42005
14. Europäische Union (2024): *Verordnung (EU) 2024/1689 über künstliche Intelligenz.* https://eur-lex.europa.eu/eli/reg/2024/1689/oj

15. OECD (2019, aktualisierte Fassung): *OECD AI Principles.* https://oecd.ai/en/ai-principles

16. Sarkar, A. (2023): *Will Code Remain a Relevant User Interface for End-User Programming with Generative AI Models?* Onward! 2023. https://www.microsoft.com/en-us/research/publication/will-code-remain-a-relevant-user-interface-for-end-user-programming-with-generative-ai-models/

17. Sarkar, A. et al. (2022): *What is it like to program with artificial intelligence?* PPIG 2022. https://www.microsoft.com/en-us/research/publication/what-is-it-like-to-program-with-artificial-intelligence/

18. Liu, M. X. et al. (2023): *What It Wants Me To Say: Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models.* CHI 2023. https://doi.org/10.1145/3544548.3580817

19. ISO/IEC (2023): *ISO/IEC 25010:2023 - Systems and software Quality Requirements and Evaluation - Product quality model.* https://www.iso.org/standard/78176.html

20. ISO/IEC/IEEE (2021): *29119-2:2021 - Software testing - Test processes.* https://www.iso.org/standard/79428.html

21. ISO/IEC/IEEE (2021): *29119-3:2021 - Software testing - Test documentation.* https://www.iso.org/standard/79429.html

22. Hollnagel, E.; Woods, D. D. (2005): *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering.* CRC Press. https://doi.org/10.1201/9781420038194

23. Lee, M. et al. (2022): *Evaluating Human-Language Model Interaction.* https://arxiv.org/abs/2212.09746