

Abschlussdiskussion

Prof. Dr. Jan Hense

KI in der Evaluation

Gemeinsame Frühjahrstagung AK Methoden und AK Verwaltung der DeGEval

21./22. Mai 2026

www.jan-hense.de | mail@jan-hense.de

Beobachtungen zur Tagung

Anwendungen Künstlicher Intelligenz in der Evaluation
Stand von Forschung und Praxis



Jan Ulrich Hense

**NUTZUNG VON KÜNSTLICHER
INTELLIGENZ IN DER EVALUATION
ÖFFENTLICHER MASSNAHMEN**

Chancen und Risiken abwägen, um hohe
Evaluationsqualität sicherzustellen

Wie bereits viele dokumentierte Anwendungsbeispiele zeigen, hat Künstliche Intelligenz (KI) das Potenzial, die Evaluationspraxis in vielerlei Hinsicht zu verändern. Eine pauschale Bewertung ihres Einsatzes fällt aufgrund verschiedener Eigenarten der Technologie schwer. Unbestreitbare Chancen sind daher sorgfältig mit den Risiken für die Evaluationsqualität abzuwägen. Auch im Angesicht von Kostendruck ist die Berücksichtigung fachlicher Standards guter Evaluation maßgeblich. Sie kann nur in professionell durchgeführten Evaluationen garantiert werden, die in eigener Verantwortung gegenstandsangemessen über den Einsatz von KI entscheiden.

Künstliche Intelligenz verändert die Evaluationspraxis

Jüngere Entwicklungen im Bereich der generativen KI erlauben heute einen breiten Einsatz auch ohne spezialisierte Kenntnisse. Besonders gilt das für *Large Language Models* (LLMs) und die zugehörigen KI-Chatbots wie zum Beispiel ChatGPT, Claude, Copilot oder Gemini, die eine natürliche Konversationschnittstelle bereitstellen. Sie können nicht nur dialogisch Fragen beantworten, sondern auch zunehmend komplexe Aufträge zu Erstellung, Bearbeitung oder Analyse von Text- und anderen Daten übernehmen. Für die Evaluation ergeben sich daraus vielfältige Anwendungsperspektiven, die im Feld bereits umfassend diskutiert werden.¹

Abb. 1: Potenzielle KI-Anwendungen im Evaluationsprozess

Phasenunspecifische Anwendungen

- Unterstützung beim Erstellen von Texten
- Exploratives Sichten, Verarbeiten und Zusammenfassen von Texten
- Ideenfindung und konzeptionelle Vorarbeiten
- Elaboration in Abwägungs- und Aushandlungsprozessen
- Programmierung unterstützen

Initiierung, Vorbereitung und Konzeption der Evaluation

- Erstellen eines Literaturüberblicks
- Neu- oder Rekonstruktion von Wirkungsmodellen
- Visualisierung von Wirkungsmodellen

Datenerhebung und -aufbereitung

- Rekonstruktion von Baseline-Daten / *Big Data* als Datenquelle
- Unterstützung der Fragebogenkonstruktion
- Automatisierung von Befragungen


¹ vgl. Hense, Jan Ulrich: Anwendungen Künstlicher Intelligenz in der Evaluation: Stand von Forschung und Praxis, PrEVal Expertise 1/2025, Frankfurt/M.



<https://preval.hsfk.de/publikationen/preval-expertisen>

Strukturierung der Anwendungsbeispiele

1. Phasenunspezifische Anwendungen
2. Initiierung, Vorbereitung und Konzeption
3. Datenerhebung und –aufbereitung
4. Datenauswertung
5. Berichterstattung und Nutzungsförderung
6. Fortlaufende und übergeordnete Aufgaben



typische
Phasen einer
Evaluation

Anwendungen

1. Phasenunspezifische Anwendungen

- Unterstützung beim Erstellen von Texten
- Sichten, Verarbeiten und Zusammenfassen von Texten
- Ideenfindung und konzeptionelle Vorarbeiten
- Elaboration in Abwägungs- und Aushandlungsprozessen
- Programmierung unterstützen

2. Initiierung, Vorbereitung und Konzeption

- Erstellen eines Literaturüberblicks
- Neu- oder Rekonstruktion von Wirkungsmodellen
- Visualisierung von Wirkungsmodellen

3. Datenerhebung und -aufbereitung

- Rekonstruktion von Baseline-Daten / Biga Data als Datenquelle
- Chatbots zur Unterstützung der Fragebogenkonstruktion
- Automatisierung von Befragungen
- Erhebung von Leistungsdaten
- Schätzung von fehlenden Daten
- Transkription von natürlicher Sprache
- Aufbereitung und Anonymisierung von Textdaten und anderen Dokumenten

4. Datenauswertung

- **Auswertung qualitativer Interviews**
- **Auswertung von Freitextantworten**
- Sentimentanalyse von qualitativen Daten
- Identifizieren von Kausalzusammenhängen / Netzwerkanalyse
- Auswertung quantitativer Daten
- Kategorisierung und Klassifizierung
- Überprüfung der Implementierungstreue

5. Berichterstattung und Nutzungsförderung

- Verschriftlichen von Befunden
- Verwendungs- und Zielgruppenspezifische Berichtsvarianten
- Visualisierungen und illustrierende Begleitmedien

6. Fortlaufende und übergeordnete Aufgaben

- Automatisierung von Rückmeldungen in Evaluationssystemen
- Wissensmanagement und Projektmanagement
- Selbstgesteuerte Weiterbildung und Capacity Development

Einige Beobachtungen

Haben wir schon begriffen, *was* GenKI alles kann?

Beispiel Vibe Coding mit Claude Code

[Beispiel Mark V. Shaney – einfacher früher Textgenerator]

[Beispiel Pollaroid – live voting tool]

MARK V. SHANEY

Character-level Markov Chain Text Generator · after Dewdney 1989

Instructions

What you can learn

AI disclaimer

Corpus

Output

Model

TRAINING CORPUS

Examples ...

Upload file

.txt .md .pdf .docx

Paste or type text here, pick an example, or upload a file ...

Look-back window (how many characters to consider): **3**

1 – chaotic

10 – close to source

Output length: **1,000 chars**

100

5,000

Seed text (first 4 chars used):

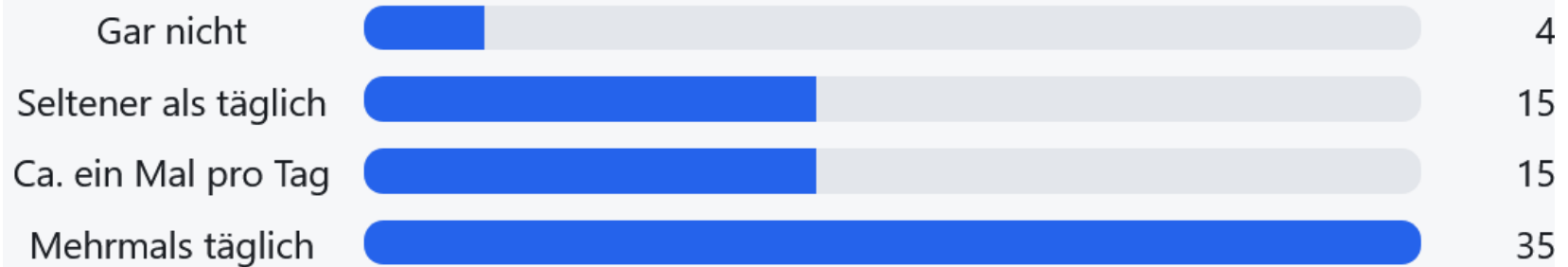
(auto-filled from corpus)

Build the Markov chain and generate text

Generate text

Wie oft nutzen Sie derzeit KI-Chatbots?

63 responses



Welche KI-Chatbots haben Sie bereits (etwas genauer) ausprobiert?

63 responses



At our finger tips

Maßgeschneiderte Datenerhebungen

Interaktives reporting

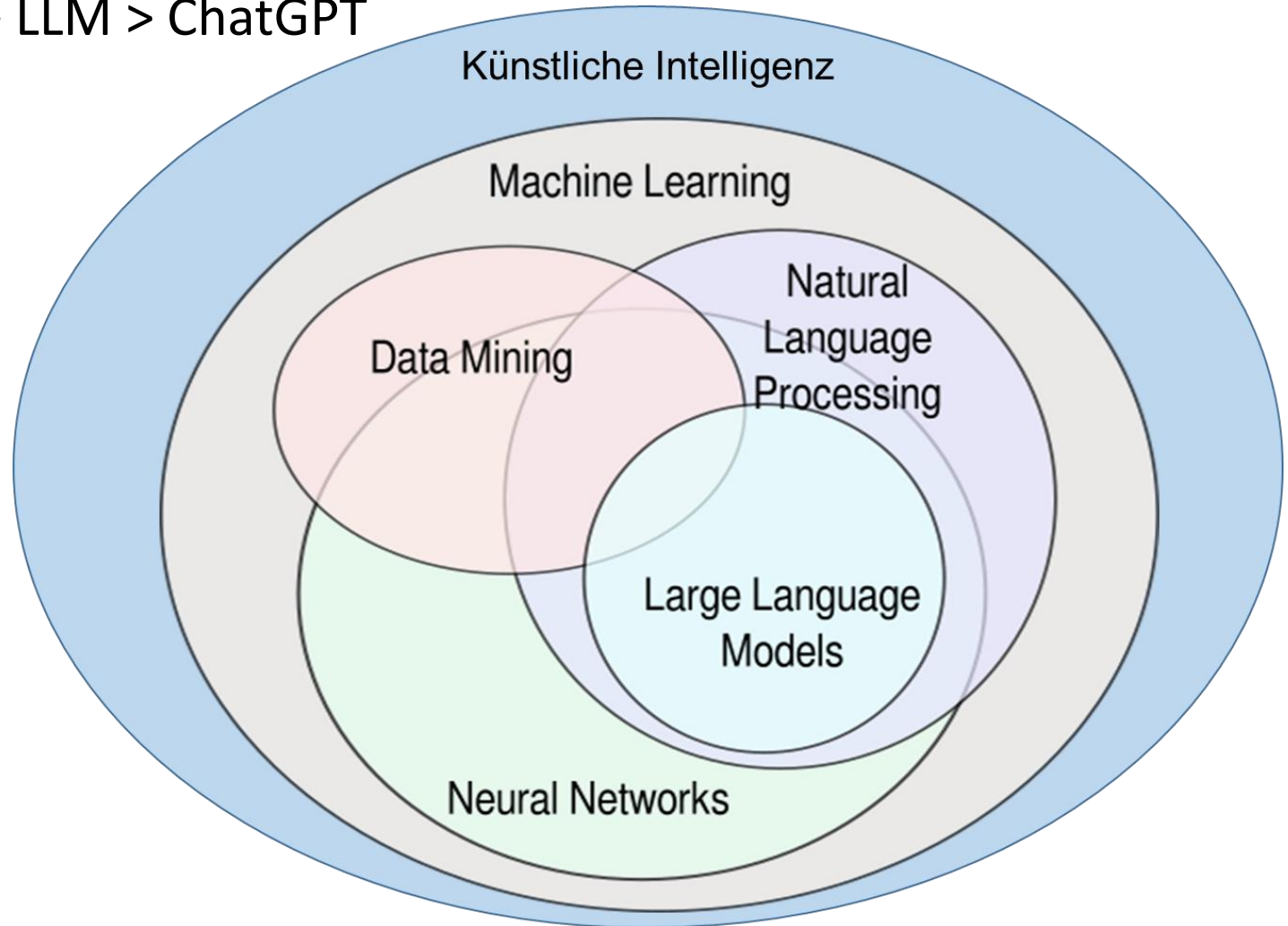
Data dashboards und data visualization

Interaktive und collaborative Wirkungsmodellierung

...

Weitere Beobachtungen

KI > GenAI > LLM > ChatGPT



Weitere Beobachtungen

Bedarf nach Orientierung führt zu Übergeneralisierungen.

- „KI führt zu...“; „KI kann/kann nicht ...“
- Magisches Denken?

Divergierenden Bewertungsperspektiven

- **Es:** Individuelle instrumentelle Nutzenevidenz / Effizienzhoffnung
- **Ich:** fachliche Kriterien
- **Über-Ich:** gesellschaftliche Kriterien

Weitere Beobachtungen

„KI-Schock“

Vom passiven Zusehen zum aktiven Gestalten

- Regelungen/Verhalten/Struktur
- „Structure shapes behavior“
- „Wo wollen wir hin?“ statt „Was macht das mit uns?“
- „Weniger als möglich ist“
- In unserer Hand: „Denke erst ich, dann die KI, oder umgekehrt?“

Evaluation von KI statt KI in der Evaluation

- Wechselmöglichkeit, Gestaltungsfähigkeit, Einfluss auf Anbietende
- Fachliche Stellungnahmen, fachinterne Regulierung

Rolle der Auftraggebenden

Evaluierende ≠ Evaluierende

Weitere Beobachtungen

Vertrauen

- Vertrauen in Menschen vs. Vertrauen in KI
- Forschungskonventionen
- Verantwortung und Haftung

Vertrauen als wichtigste Währung der KI?

- Ethical codes z. B. Anthropic constitution
- Amanda Askell: AI Alignment
- Transparenz durch Offenlegung von Systeminterna

System Prompts

 Copy page 

See updates to the core system prompts on claude.ai and the Claude [iOS](#) and [Android](#) apps.

Claude's web interface (claude.ai) and mobile apps use a system prompt to provide up-to-date information, such as the current date, to Claude at the start of every conversation. The system prompt also encourages certain behaviors, such as always providing code snippets in Markdown. This prompt is periodically updated to improve Claude's responses. These system prompt updates do not apply to the Claude API. Updates between versions are bolded.

Claude Opus 4.7

> **April 16, 2026**

Claude Sonnet 4.6

> **February 17, 2026**

<https://platform.claude.com/docs/en/release-notes/system-prompts>

Claude Opus 4.6

Claude Opus 4.7 System prompt

<claude_behavior> <product_information> Here is some **information about Claude** and Anthropic's products in case the person asks: This iteration of Claude is Claude Opus 4.7 from the Claude 4.7 model family. The Claude 4.7 family currently consists of Claude Opus 4.7. Claude Opus 4.7 is the most advanced and intelligent model.

...

If the conversation feels risky or off, Claude understands that saying less and giving shorter replies is safer for the user and runs less risk of **causing potential harm**.

...

If a user indicates they are ready to end the conversation, Claude does not request that the user stay in the interaction or try to elicit another turn and instead **respects the user's request to stop**.

Verstehen Sie, wie ein KI-Chatbot sein Output erzeugt?

62 responses




Wissen Sie, was ein System Prompt ist und könnten erklären, was er tut?

61 responses



(post hoc Zusammenfassung von Claude Opus 4.7)

Kategorie (Originalbegriffe)	Anzahl
Trainingsdaten (Trainingsdaten, Trainingsmaterial, Daten, Datensatz, Datenquelle, Hinterlegte Daten, Vorliegende Daten, Material, Infos, mit der die KI trainiert wurde)	23
Temperatur (Temperatur, Temperature, Temperatureinstellung, Einstellung der Temperatur)	11
Modell/Version (Modell, Modell des Chatbots, Modell Version, Das gewählte Modell, Modellgröße, Bezahlversion, Abo Modell, Free oder Bezahlversion, Modell (Bezahlt oder „free“), Version - kostenfrei oder nicht, paywall, Art des Zugangs)	11
Chatverlauf/vorherige Eingaben (Chatverlauf, Bisherige Eingaben, Vorangegangener Input, Vorherige Anfragen, vorherige Chats, Vorheriger Chat, die bereits vorher gestellt worden, Projektspeicher)	9
Sprache (Sprache, Sprache selbst, Sprachstil, Sprach Niveau, Grammatik)	7
Input/Prompt-Eigenschaften (Input, Mein Prompt, Die Frage selbst, Fragen, Aufgabe, Thema, Reichweite der Frage, Größe des Inputs, Antwortlänge, Rolle, Verwandte Themen)	11
Systemprompt (Systemprompt, Setting des System Chats, Vorgaben und Einstellungen)	6
Datenqualität (Datenqualität, Qualität der Daten, Qualität der Inputs, Vorabqualitätskontrolle, Genauigkeit)	5
Programmierung/Algorithmen (Programmierung, Algorithmen, Algorithmus, die Programmierung des Chatbots, Technologie, Technologie-Entwicklungsprozesse)	6
Training allgemein (Training, Pretraining, Trainingsdauer, Forschung)	5
Wahrscheinlichkeit (Probability, Wahrscheinlichkeit, Wahrscheinlichkeiten, Häufigkeiten)	4
Daten anderer Nutzer (bisherige Eingaben von anderen Usern, die bereits in anderen Chats eingegeben wurden, bestärkende Antworten)	3
Zeit/Aktualität (Zeit, Aktualität, Verfügbarkeit)	3
Hochgeladene Dokumente (die eingegebenen Dokumente, die wir dazu hochladen, Quellen) 	3

Die Karten werden neu gemischt: Anonymisierung und Pseudonymisierung als Beispiel

Datenschutzproblematik bei Übertragung von qualitativen Daten an Online-Dienste

Selbst bei vollständiger Anonymisierung kann u.U. auf Personen, Institutionen und Orte rückgeschlossen werden.

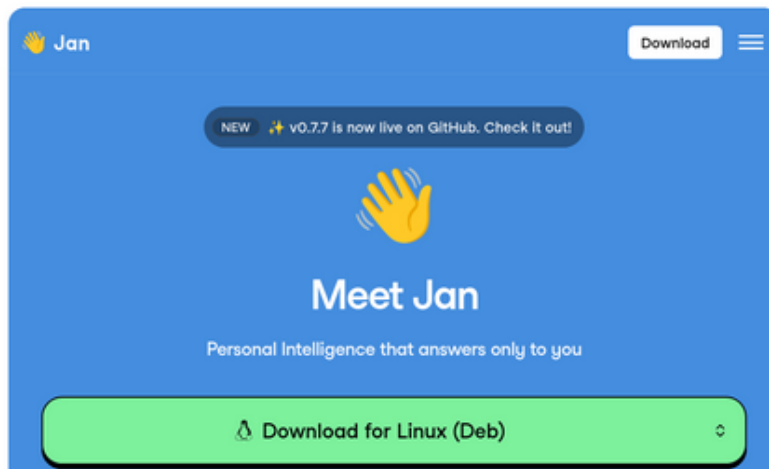
Dresing et al. (submitted). Datenmaskierung mittels Pseudo- und Anonymisierung reicht nicht! Methodische, ethische und datenschutzrechtliche Herausforderungen bei der KI-gestützten Verarbeitung personenbezogener qualitativer Daten. *Zeitschrift für Evaluation*.

Lokale KI-Modelle als Lösung!?

WELCOME TO JAN ET AL.

Lokal installierte Anwendungen könnten einige Probleme des KI-Einsatzes in Evaluationen lösen. Die Betonung liegt leider meistens immer noch auf dem Konjunktiv.

Von Jan Hense am Mi., 11.02.2026 - 20:23



Ein Aufruf von Kai Dröge, dem Autor der Transkriptionssoftware [noScribe](#) in der Mailingliste [QSF-L](#), den ich hiermit gerne umsetze und weiterverbreite (s.u.), bietet den willkommenen Anlass, etwas zum Thema **lokal installierbare KI-Modelle** zu schreiben.

Lokale KI-Modelle hatte ich schon in meiner

<https://www.jan-hense.de/blog/welcome-jan-et-al>

Ein Schritt zurück

Die historische Perspektive

18. Jhd.

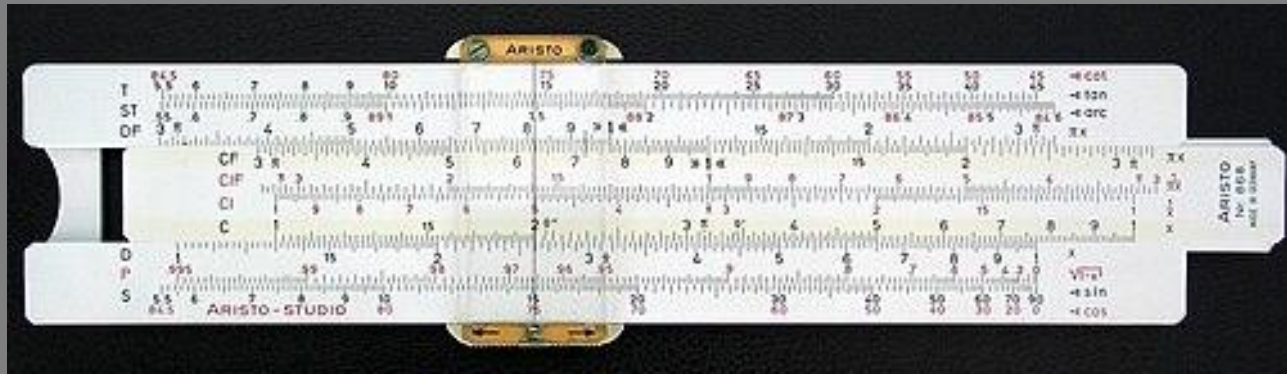
- Aufklärung
- Ideengeschichtliche Voraussetzungen

19. Jhd.

- Statistik und Wahrscheinlichkeitstheorie
- Empirische Sozialforschung
- Experimentieren als wissenschaftliche Methode (z. B. Wilhelm Wundt)

1930er

- „Marrienthal Studie“ (Marie Jahoda, Paul Lazarsfeld):
- Methodenkombination, Feldstudie, qualitative Sozialforschung
- Erfindung der „Evaluation“ im heutigen Sinne
- *Eight year study* (Ralph. W. Tyler)



https://commons.wikimedia.org/wiki/File:Sliderule_2005.jpg

https://commons.wikimedia.org/wiki/File:Rechenmaschine_Walther_WSR_160.JPG

[https://commons.wikimedia.org/wiki/File:Continental-Schreibmaschine_mit_Tabulator_\(2021-02-21_Sp_04\),_re_\(2\).jpg](https://commons.wikimedia.org/wiki/File:Continental-Schreibmaschine_mit_Tabulator_(2021-02-21_Sp_04),_re_(2).jpg)

<https://creativecommons.org/licenses/by-sa/3.0/>

Technologische Revolutionen in der Evaluation

1950er Tonbandgeräte

- Vereinfachte Nutzung qualitativer Daten

1960er Mainframe Computer

- Auswertung großer quantitativer Datensätze
- „Landmark“ Evaluationsstudien z.B. *The First Year of Sesame Street: An Evaluation* (Ball & Bogatz)
- *Campbell & Stanley (1963): Experimental and Quasi-Experimental Designs for Research*
- *Suchman (1969). Evaluative Research*
- Erste Evaluationsgesellschaften (ERS 1976, ENet 1978)

Technologische Revolutionen in der Evaluation

1980er PC

- Evaluation als unabhängige Dienstleistung
- Von rein wissenschaftlicher Tätigkeit zur Marktdynamik
- „Age of Expansion“: Notwendigkeit zur Professionalisierung
- Gründung der AEA 1986
- Erste Evaluationsstandards (JCSEE, 1981)

1990er Internet & 2000er Web 2.0

- Beschleunigte Kommunikation
- Vereinfachter Informationszugriff
- Nationale Evaluationsgesellschaften
- Communities of Practice
- Neue Formen der Datenerhebung & neue Arten von Daten

Technologische Revolutionen in der Evaluation

2020er

- KI betritt das Spielfeld
- und jetzt ?

YouTube short:
Elle Cordova – „Inventions hanging out“
www.youtube.com/shorts/NHat7IaZ488

Parallele Entwicklungen

Evaluationstheorie:

- Entwicklung von Evaluation als Transdisziplin

Evaluationspraxis:

- von elitärer Praxis einzelner zum „anyone can do it“

Spannungsfeld zwischen
Professionalisierung und Deprofessionalisierung



Alles andere als stabil: Zum Stand der Professionalisierung

- ✓ Evaluationsstandards
- ✓ Fachgesellschaften
- ✓ Fachzeitschriften
- ✓ Forschung über Evaluation & professionelle Wissensbasis

- ✗ keinerlei Schließungsmechanismen
- ✗ Ungeschützte Berufsbezeichnung
- ✗ keinerlei Sanktionsmöglichkeiten gegen schlechte Praxis
- ✗ Marktdruck erzeugt Verführbarkeit
- ✗ Umfeld: gesellschaftlicher Trend zum Postfaktischen
- ✗ Intern: Unklarheit über den professionellen Kern

Brauchen wir das noch oder kann das schon weg?

Artificial Intelligence and the Future of Evaluation: From Augmented to Automated Evaluation

STEVE JACOB, Université Laval, Quebec, Canada

The recent developments in artificial intelligence (AI) are revolutionizing professional practices across various professional fields, including evaluation. With its advanced automation and learning capabilities, AI is bringing significant changing to the way organizations and societies function. Evaluation is no exception to this trend, even though evaluators are adopting AI at a slower pace. This article examines ongoing applications that already improve and enhance the evaluation practice. We advance our discussion by exploring the potential impact of AI on the policy cycle. Subsequently, we analyze the potential **incorporation of evaluation into autonomous AI systems that could design, implement, and evaluate public policies with minimal to no human supervision.**

CCS Concepts: • **Social and professional topics** → **Professional topics**; • **Computing methodologies** → **Artificial intelligence**;

Additional Key Words and Phrases: Public policy evaluation, artificial intelligence, skills and competencies, generative AI

ACM Reference Format:

Steve Jacob. 2025. Artificial Intelligence and the Future of Evaluation: From Augmented to Automated Evaluation. *Digit. Gov. Res. Pract.* 6, 1, Article 10 (February 2025), 10 pages. <https://doi.org/10.1145/3696009>

Was ist der professionelle Kern der Evaluation und warum ist er wichtig?

Abbott (1988): Hoheitsgebiet einer Profession liegt im nicht delegierbaren Schluss von Diagnose auf Maßnahme

Diagnose → Inferenz → Maßnahme

Wo und wann findet diese Inferenz bei der Evaluation statt?

Ausgangslage → Evaluationsdesign (inception report etc.)

Daten → Werturteil (“erfolgreich”, “verbesserungswürdig”, ...)

Interpretation → Empfehlungen

Finger weg, KI?

- Entwicklung von Evaluationsdesign
- Bewerten im eigentlichen Sinne (Logik der Evaluation)
 - Kriterien: Woran lesen wir ob, ob etwas „gut“ ist?
 - Vergleichswerte: ab wann ist es „gut“?
 - Synthese: Wie wägen wir divergierende Kriterien gegeneinander ab?
- Ableiten von Verbesserungen und Empfehlungen

Also wäre alles andere delegierbar? An die KI? Z. B.

- Kontextanalyse?
- Datenerhebung?
- Datenauswertung?

It all comes down to **Bewertungskompetenz**

Verlass' dich nicht auf den Autopilot:
Wir dürfen nicht vergessen, wie man es selber macht.

Bewertung der Ergebnisse

- Wie gut ist das, was rauskommt?

Bewertung des Prozesses

- Wurde es auf die richtige Art und Weise gewonnen?
- Wie ist die Seilbahn auf den Berg gekommen?

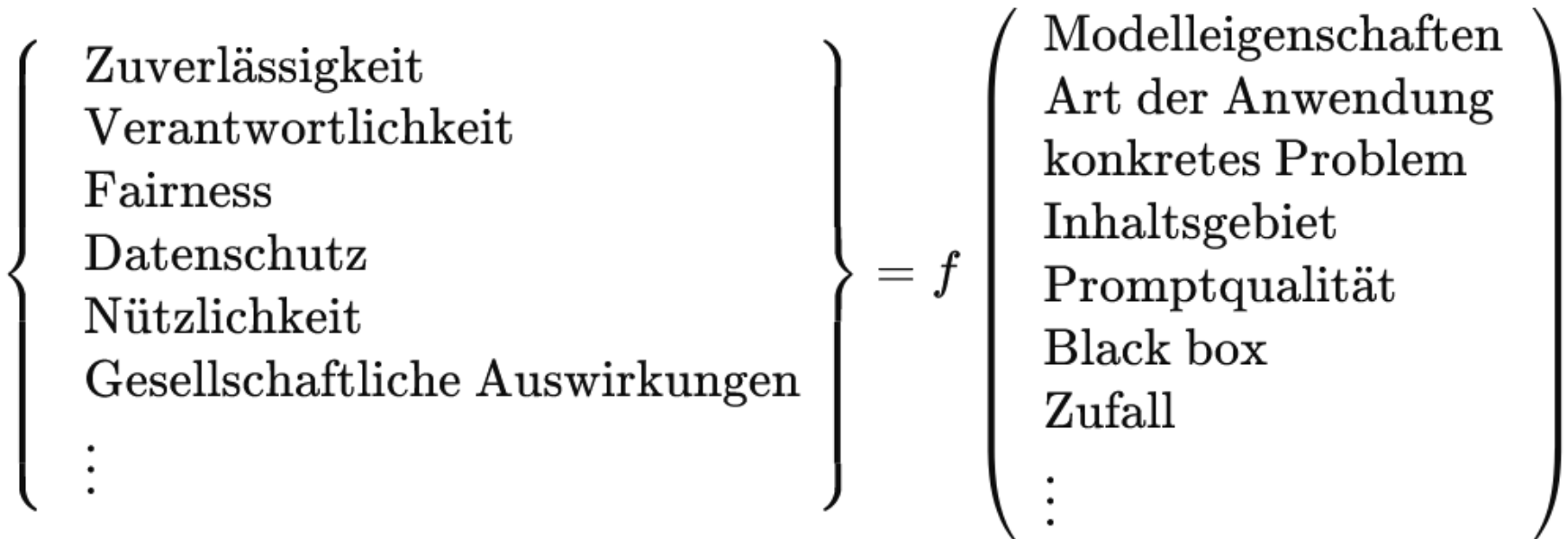
Müssen wir „on top“ verstehen können, wie generative KI arbeitet?

Unterschiede zu herkömmlichen Technologien

1. Polyvalenz
2. Stochastische Funktionsweise
3. Unsichere Reproduzierbarkeit
4. Koproduktion der Ergebnisqualität
5. Diffuse Systemgrenzen
6. Einbettung
7. Wechselseitige Anpassung
8. Volatilität

Frühere technologische Neuerungen waren immer deterministisch.
GenAI/LLMs sind es nicht

Bewertung von KI



Auswirkungen auf die Evaluation?

Welche Evaluation?

Mit “Evaluation“ meinen wir vier unterschiedliche Dinge:

1. Systematisches Bewerten
2. Konkrete Evaluationsstudien
3. Das, was Evaluierende tun (Evaluationsdienstleistung)
4. Ein sich professionalisierendes Tätigkeitsfeld

Wo **werden** welche wir Auswirkungen sehen?

Wo **wollen** wir welche Auswirkungen zulassen?



Abschlussdiskussion

Wird KI alles in allem Evaluation besser oder schlechter machen?

62 responses

