

Arbeitskreis
Methoden
in der Evaluation



Arbeitskreis
Professionalisierung



„Alles ChatGPT oder was? – Chancen und Herausforderungen Künstlicher Intelligenz (KI) in der Evaluation“

*Session A4: Gemeinsame Session der Arbeitskreise
Methoden und Professionalisierung*

Ablauf der Session

Intro (5 min)

Dr. Alexander Kocks, Dr. Franziska Heinze
& Jessica Prigge

Kurzinput „Künstliche Intelligenz in der
Evaluierungspraxis - Von der
technologischen Revolution in den
Arbeitsalltag“ (20 min)

Kai Rompczyk, DEval

Diskussionsgruppen (30 min)

A) Chancen und Potenziale

B) Herausforderungen und
Lösungsansätze

C) Professionalisierungsanforderungen

Zusammenfassung und Ausblick (30 min)

Künstliche Intelligenz in der Evaluierungspraxis

Von der technologischen Revolution in den Arbeitsalltag

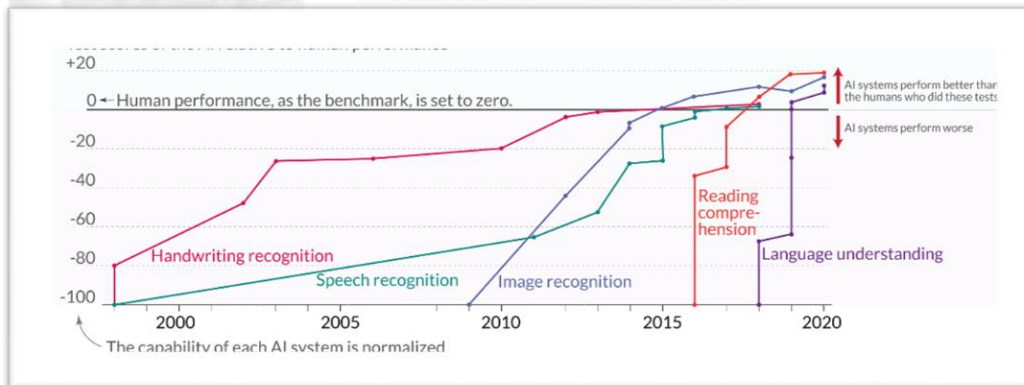
DeGEval Jahrestagung – Magdeburg

Kai Rompczyk (DEval)

Donnerstag, 14.09.08.2023

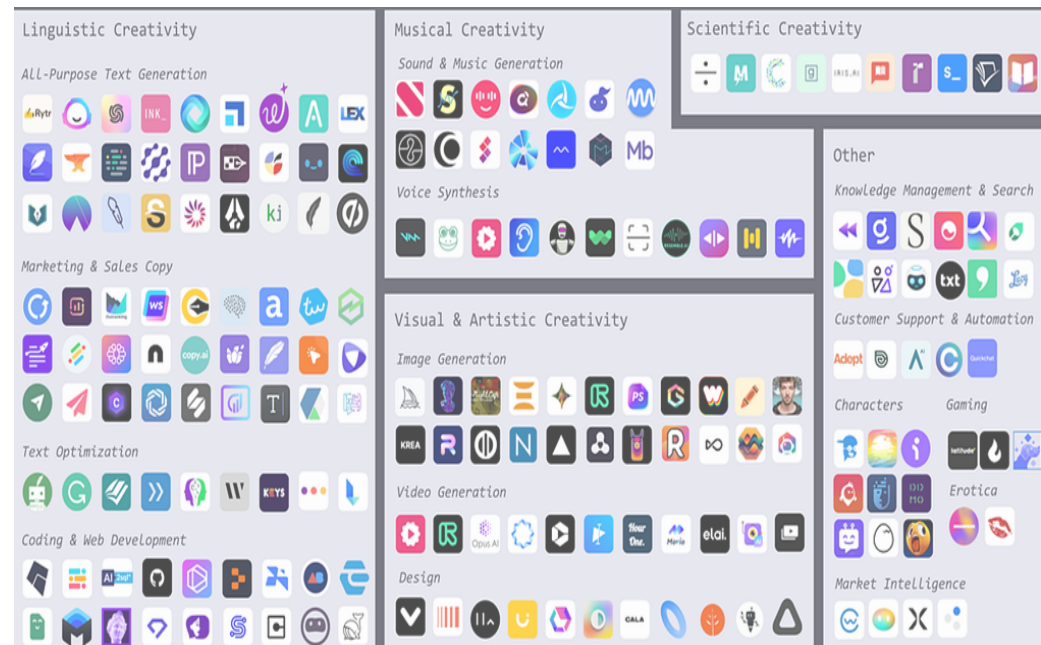
Technological Revolution with an unpredictable future

- 1) Invention of the principle of **autonomous learning** and **acting systems**
- 2) Exponential increase in **computing power** and **data (sources)**
- 3) AI capabilities are becoming more **generalist** and **multimodal**
- 4) **Shrinking barriers and costs** for the use of the new technology
- 5) „AI systems are now approaching **human level performance** on a large variety of tasks” (Fei et al. 2022)



Anwendungsbeispiele für die Evaluierungspraxis

- 1) Evidenzaufbereitung
- 2) Analysen aus multiplen Perspektiven
- 3) Qualitätssicherung und Peer Review
- 4) Internes Wissensmanagement
- 5) Automatisierte Datenaufbereitung und –analyse



Anwendungsbeispiele

1) Portfolioanalyse und Evidenzaufbereitung: Thematische Schwerpunkte aus Evaluierungsberichten in Burkina Faso



- Ein bedeutender Fokus lag auf der Gesundheit und dem Lebensstandard der Menschen, insbesondere im Hinblick auf die Risiken durch offene Abwasserkanäle und Umweltschäden in bestimmten Gebieten.
- Es gab Bemühungen, die Leistungsfähigkeit der Gemeinden zu verbessern. Trotz einiger Beiträge zur Erreichung der Ziele gab es Herausforderungen, insbesondere aufgrund schlechter Governance auf kommunaler Ebene und Schwächen in der Programmausführung.
- Projekte zielten darauf ab, den Menschen in den Programmregionen nachhaltigen Zugang zu sozialer und wirtschaftlicher Infrastruktur sowie zu politischen Beteiligungsformen zu ermöglichen.
- Es wurden Maßnahmen zur Einkommensgenerierung durchgeführt, darunter die

2) Analysen und Bewertungen aus multiplen Perspektiven



"Während die Veranstaltung zweifellos eine akademische Tiefe verspricht und die Spannungen zwischen wissenschaftlicher Evidenz und politischem Handlungsdruck beleuchtet, fehlt ihr der kritische Blick auf die zugrunde liegenden Klassenstrukturen. Es besteht die Gefahr, dass sie sich in theoretischen Diskussionen verliert, ohne die realen Machtverhältnisse und Klassenkämpfe, die solche Spannungen überhaupt erst erzeugen, anzusprechen."



„Das Grußwort legt zwar Wert auf rasches Handeln, was ich schätze, aber die Veranstaltung scheint sich zu sehr auf akademische Diskurse und wissenschaftliche Genauigkeit zu konzentrieren. In der echten Welt brauchen wir weniger Theorie und mehr praktische Lösungen, die funktionieren.“

Bewertung des Grußwortes der DeGEval Jahrestagung 2023, erstellt von ChatGPT

3) Qualitätssicherung und Peer Review

„Nach Durchsicht der Dokumentauszüge fand ich mehrere Textpassagen, die auf die Verwendung von Evaluierungsmethoden in dem Bericht hinweisen.

(...)

Das Dokument erwähnt, dass in der Evaluierung eine Regressionsanalyse verwendet wurde, um den Einfluss von Konflikten auf die Wirksamkeit von Anpassungsmaßnahmen zu untersuchen. Der Text besagt, dass *„zum Thema Konfliktsensibilität nutzte sie Daten zur Erfolgsbewertung von Anpassungsmaßnahmen und untersuchte den Einfluss von Konflikten auf die Wirksamkeit von Anpassungsmaßnahmen anhand von Regressionsanalysen“*. (...)

(...)

Als Prüfer bin ich überzeugt, dass in der Evaluierung angemessene Methoden verwendet wurden, um die Anpassung der deutschen Entwicklungszusammenarbeit an den Klimawandel zu beurteilen.“

4) Internes Wissensmanagement: Dokumenten-Chatbot



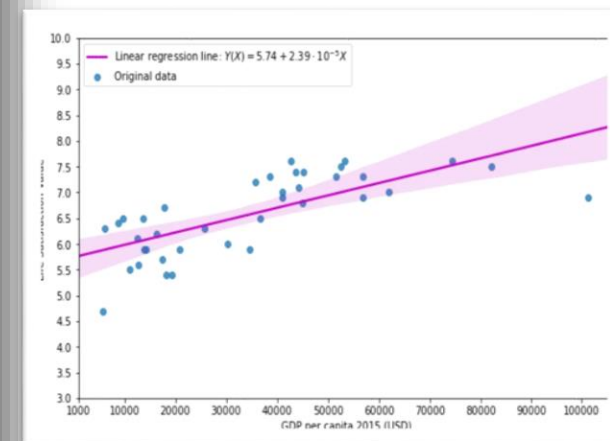
Die Erhebungsphase ist ein zentraler Bestandteil jeder Evaluation. Hier sind fünf Stichpunkte aus den DeGEval-Standards, die während dieser Phase besonders zu beachten sind:

1. **Auswahl und Umfang der Informationen:** Die Auswahl und der Umfang der gesammelten Informationen sollten so gestaltet sein, dass sie eine angemessene Beantwortung der zu untersuchenden Fragestellungen ermöglichen und die Informationsbedürfnisse der Auftraggeber sowie weiterer Beteiligter berücksichtigen.
2. **Transparenz von Werthaltungen:** Werthaltungen der Beteiligten und Betroffenen, die ihre Perspektiven und Annahmen beeinflussen und einen Einfluss auf die Evaluation und Interpretation ihrer Ergebnisse haben, sollten transparent dokumentiert werden.
3. **Angabe von Informationsquellen:** Die in einer Evaluation verwendeten Informationsquellen sollten genau dokumentiert werden, um die Zuverlässigkeit und Angemessenheit der Informationen einschätzen zu können.
4. **Valide und reliable Informationen:** Die Erhebungsverfahren und Datenquellen sollten so ausgewählt werden, dass die Zuverlässigkeit der Daten und ihre Gültigkeit in Bezug auf die Beantwortung der Evaluationsfragestellungen nach fachlichen Maßstäben gewährleistet sind.
5. **Systematische Fehlerprüfung:** Die in einer Evaluation gesammelten und verarbeiteten Informationen sollten systematisch auf Fehler überprüft werden.

5) Automatisierte Datenaufbereitung und -analyse

Hier sind die Ergebnisse der linearen Regression:

1. Das Modell erklärt etwa 48,9% der Varianz in der abhängigen Variable ``read`` (R-quadrat = 0,489).
2. Die Koeffizienten für die unabhängigen Variablen sind wie folgt:
 - **write**: Für jeden zusätzlichen Punkt in der Schreibbewertung wird ein Anstieg von etwa 0,441 Punkten in der Lesebewertung erwartet, bei sonst gleichen Bedingungen.
 - **female**: Weibliche Schüler haben im Durchschnitt eine um 2,706 Punkte niedrigere Lesebewertung als männliche Schüler, bei sonst gleichen Bedingungen.
 - **math**: Für jeden zusätzlichen Punkt in der Mathematikbewertung wird ein Anstieg von etwa 0,321 Punkten in der Lesebewertung erwartet, bei sonst gleichen Bedingungen.
 - **prog (allgemein gegenüber Berufsbildung)**: Schüler im allgemeinen Programm haben im Durchschnitt eine um 0,518 Punkte höhere Lesebewertung als Schüler im Berufsbildungsprogramm, bei sonst gleichen Bedingungen. Dieser Effekt ist jedoch statistisch nicht signifikant ($p > 0,05$).
 - **prog (akademisch gegenüber Berufsbildung)**: Schüler im akademischen Programm haben im Durchschnitt eine um 1,811 Punkte höhere Lesebewertung als Schüler im Berufsbildungsprogramm, bei sonst gleichen Bedingungen. Auch dieser Effekt ist statistisch nicht signifikant ($p > 0,05$).



Advanced Data Analysis Beta

Possibilities of future use of AI in the evaluation phases



1 Concept, Inception

- Participatory communication tools
- Preliminary analyses on a larger scale



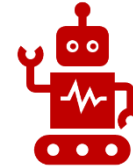
2 Data collection

- Translation
- Transcription
- Interview chatbots
- Big data scraping
- Data preparation
- Anonymization and data security



3 Analysis

- Organizing and summarizing data and documents
- Performing multiple and diverse analyses (qualitative + quantitative, deductive + inductive)
- Prediction
- Providing alternative interpretations



4 Synthesis and reporting

- Evidence synthesis
- Quality assurance
- Standardized reporting
- bibliography creation



5 Dissemination

- Individualized, interactive presentation of reports
- (dashboards, chatbots, videos etc.)

Risiken

- 1) **Big Data vs. Datenschutz:** sind gängige Verfahren des Umgangs mit sensiblen Daten noch ausreichend?
- 2) **Verzerrungen und Vorurteile** auf Grundlage der Trainingsdaten (Bias)
- 3) **Intransparenz** durch **Blackbox** Künstlicher Intelligenzen
- 4) Scheinbar menschliche Kompetenzen mit **nichtmenschlichen Fehlern** (u.a. eingeschränkte Logikkompetenzen, Halluzinationen, kein Zugang zur realen Welt)
- 5) Abgabe von **Verantwortung und Kompetenz** durch Automatisierung
- 6) **Erosion des Vertrauens** in etablierte Wissensnetzwerke durch KI-Einsatz (u.a. konkurrierende KIs, Deepfakes)
- 7) **Sozialökologische Nachhaltigkeit** (Probleme mit Urheberschaft des Trainingsmaterials, Entwicklung der KI durch Ausbeutung in Niedriglohnländern, CO²-Abdruck)
- 8) Herausforderungen für **Evaluierungsstandards** (z.B. Nachvollziehbarkeit, Transparenz, Unabhängigkeit)

Vielen Dank für Ihre Aufmerksamkeit!

Weiterführende Literatur

Autor, D. (2022): The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty, [Autor-Brookings-Tech-Inequality-NBER-rev-20220726.pdf](#)

Autor, D.H., Levy, F., Murnane, R.J. (2003): The skill content of recent technological change: an empirical exploration. Q. J. Econ. 118(4), 1279–1333

Bang, Y et al. (2023): A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, [2302.04023.pdf \(arxiv.org\)](#)

Bubeck, S. (2023): Sparks of Artificial General Intelligence: Early experiments with GPT-4, [2303.12712.pdf \(arxiv.org\)](#)

Connachie, M. (2023): [Exploring the opportunities and risks of ChatGPT: Highlights from the MERL Tech Natural Language Processing \(NLP-COP\) Meeting - MERL Tech](#)

Davies, R. (2023): Exploring the text analytic capacities of GPT4, [Exploring the text analytic capacities of GPT4 – Monitoring and Evaluation NEWS \(mande.co.uk\)](#)

DEval (2018): Methods and Standards 2018, STANDARDS FOR DEVAL EVALUATIONS, [DEval_Policy_Brief_Methods_Standards_2018_EN.pdf](#)

European Commission (2019), Ethics Guidelines for Trustworthy AI, [Ethics Guidelines for Trustworthy AI | FUTURIUM | European Commission \(europa.eu\)](#)

European Commission (2021): Regulation of the European Parliament And of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, [resource.html \(europa.eu\)](#)

Gilardi (2023): ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks, [2303.15056.pdf \(arxiv.org\)](#)

Heine, M. et al. (2023), *Künstliche Intelligenz in öffentlichen Verwaltungen: Grundlagen, Chancen, Herausforderungen und Einsatzszenarien*, Edition eGov-Campus, Springer Fachmedien, Wiesbaden, doi:10.1007/978-3-658-40101-6.

Heine, M. et al. (2023), *Künstliche Intelligenz in öffentlichen Verwaltungen: Grundlagen, Chancen, Herausforderungen und Einsatzszenarien*, Edition eGov-Campus, Springer Fachmedien, Wiesbaden, doi:10.1007/978-3-658-40101-6.

OpenAI (2023): ChatGPT, the Start of a New Era A Bright and Gloomy Future, [gpt-4.pdf \(openai.com\)](#)

OECD (2023): Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading, Educational Research and Innovation, OECD Publishing, Paris, <https://doi.org/10.1787/73105f99-en>.

Raimondo et al. 2023, [Setting up Experiments to Test GPT for Evaluation | Independent Evaluation Group \(worldbankgroup.org\)](#)

Reed et al. (2022): A Generalist Agent, [\(openreview.net\)](#)

Rozado, D. (2023), „The Political Biases of ChatGPT“, *Social Sciences*, Multidisciplinary Digital Publishing Institute, Vol. 12, Nr. 3, S. 148.

Sezgin (2023): ChatGPT as a Game-changer Tool for Qualitative Research with Smart Prompts, [\(PDF\) ChatGPT as a Game-changer Tool for Qualitative Research with Smart Prompts \(researchgate.net\)](#)

Sun, F. (2022): ChatGPT, the Start of a New Era. A Bright and Gloomy Future , [ChatGPT_the_Start_of_a_New_Era.pdf \(feisun.org\)](#)

Safadi, Hans (2023): GPT-3 as your Qualitative Research Assistant, [PowerPoint Presentation \(uga.edu\)](#) – GPT-3 as Qualitative Research Assistant

Tiunova, A. und F. Muñoz (2023), „Chatgpt: Using Ai in Social Studies Academic Research“, *Available at SSRN 4451612*.

United States Government or the European Union (2022): The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America, [The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America \(europa.eu\)](#)

Typologie von Tätigkeiten im Hinblick auf Automatisierbarkeit

- Analytische und kognitive Tätigkeiten
- Interaktive Tätigkeiten
- Manuelle Tätigkeiten

- Routine
- Nicht-Routine

Tabelle 1: Fünf Tasks-Dimensionen

Analytische Nicht-Routine-Tasks	Forschen, analysieren, evaluieren, planen, konstruieren, designen, entwerfen, Regeln/Vorschriften ausarbeiten, Regeln anwenden und interpretieren
Interaktive Nicht-Routine-Tasks	Verhandeln, Interessen vertreten, koordinieren, organisieren, lehren oder trainieren, verkaufen, einkaufen, Kunden werben, werben, unterhalten, präsentieren, Personal beschäftigen oder managen
Kognitive Routine-Tasks	Kalkulieren, Buchhaltung machen, Texte/Daten korrigieren, Länge/Höhe/Temperatur messen
Manuelle Routine-Tasks	Maschinen bedienen oder kontrollieren, Maschinen ausstatten
Manuelle Nicht-Routine-Tasks	Reparieren oder renovieren von Häusern/Wohnungen/Maschinen/Fahrzeugen, restaurieren von Kunst/Denkmälern, Gäste bedienen oder beherbergen

Dengler et al. 2014: Berufliche Tasks auf dem deutschen Arbeitsmarkt
https://doku.iab.de/fdz/reporte/2014/MR_12-14.pdf

What can ChatGPT do for Monitoring and Evaluation?

Summarising existing text. This can include condensing text or pulling out key points, findings or headlines. Evaluators often have to do this task during the data analysis stage when summarising information from documents or interview transcripts. At the report writing stage, it could also be used for drafting executive summaries and abstracts.

Comparing multiple texts. ChatGPT can be asked to compare and identify the differences between texts. It can also be asked to rate aspects like which text is more optimistic, likely or realistic. This could be useful for comparing/triangulating different evidence sources.

Analysing individual texts. This covers a wide range of capabilities from (1) information extraction (actor/place/relationship/statistics); (2) text classification similar to qualitative thematic analysis covering both deductive and inductive coding (including sentiment analysis); and (3) causal relationship extraction, the process of identifying and synthesising causal claims.

Analysing tables and images. ChatGPT can write a supporting narrative to data tables and also images. This could be very useful during both the analysis and reporting evaluation stages.

Creating (generating) text. ChatGPT can draft text based on the information it already has in its model (from the internet in 2021). This could be potentially useful for drafting text for an introduction or a context section of an evaluation report. ChatGPT could also be asked to draft text based on text that is shared with it; e.g., drafting a results or discussion section. ChatGPT is very limited with the amount of text that can be uploaded but other options are now being developed (see below).

Connachie (2023)

Austausch- und Diskussionsfragen:

1. Welche Potentiale und Chancen birgt KI für und in Evaluationen? Inwiefern sind welche Evaluationsstandards davon berührt?
2. Welche Herausforderungen bringt KI für und in Evaluationen mit sich? Inwiefern sind welche Evaluationsstandards davon berührt? Welche Lösungen gibt oder braucht es?
3. Welche Veränderung hinsichtlich der Professionalisierungs-anforderungen an Evaluierende stellen sich angesichts von Entwicklungen im Bereich KI?

Dokumentation der Ergebnisse

Nachdem in Kleingruppen Chancen, Risiken und Herausforderungen sowie Folgen für die Evaluation als Profession besprochen wurden, wurden zentrale Aspekte gemeinsam diskutiert. Folgend dokumentieren wir diese entlang der Notizen, die auf Kärtchen von den Teilnehmenden festgehalten und von uns gruppiert wurden.

Welche **Potenziale und Chancen** birgt KI für und in Evaluationen? Inwiefern sind welche Evaluationsstandards davon berührt?

Einen deutlichen Schwerpunkt in der Diskussion bildeten Aspekte, die **Nützlichkeits-Standards** betreffen:

- kann Berichterstattung unterstützen:
 - Kurzfassung/Zusammenfassung schreiben lassen
 - Übersetzung in andere Sprache (z.B. Englisch)
 - Sprachtyp Zielgruppenbezogen adaptieren (z.B. für Kinder)
 - Routine-Passagen in Texten schreiben

(Fortsetzung)

- Lektorat
- Entwurfsskizzen schreiben
- Textgewinnung
- Datenanalyse
- Syntaxen erstellen lassen
- Skripte schreiben
- Fehlermeldungen erklären
- Analyse größerer Fallzahlen in der qualitativen Forschung (ggf. sogar Vollerhebung)
- API direkt speichert keine Daten oder „nur“ für 30 Tage
- Überprüfung Einhaltung der DeGEval-Standards mithilfe von KI?

(Fortsetzung)

- Angebot schreiben lassen
- Höhere Geschwindigkeit der Bearbeitung
- Vollständigkeit der Berichterstattung (N6)
- Rechtzeitigkeit (N7)

Weitere diskutierte Potenziale und Chancen beziehen sich auf **Genauigkeits-Standards:**

- KI als Gesprächs-/Reflexionspartner, um verschiedene Perspektiven einzufangen (Multiperspektivität)
- Systematische Fehlerprüfung (G6)
- Korrekturschleifen anfordern (G)

Auch in Bezug auf **Durchführbarkeits- und Fairness-Standards** werden einige Vorteile gesehen:

- Reiz des Kostenvorteils (D)
- Effizienz (D3): schneller, leichter zugänglich
- Automatisierte Verarbeitung von qualitativen Daten
- Mehrwert für Primärdatenerhebung
- Jedes Ministerium wird eigene KI-Anwendung erarbeiten (Mehrwert in der feldspezifischen Anreicherung)
- Formale Vereinbarungen (F1)
- Schutz individueller Rechte (F2; Anonymisierung von Texten)
- Zugänglichmachen für Berichte von Betroffenen (F5)

Welche **Herausforderungen** bringt KI für und in Evaluationen mit sich? Inwiefern sind welche Evaluationsstandards davon berührt?

Ethische und rechtlich-datenschutzbezogene Risiken, die sich besonders auf **Fairness-Standards** beziehen, bildeten einen Schwerpunkt:

- Wann braucht es Menschen?
- Fehleranfällige Evaluator*in bleibt im Mittelpunkt
- Unparteilichkeit durch Bias gefährdet (betrifft allerdings auch uns Menschen)
- Inwiefern Verstöße gegen DSGVO? Schutz individueller Rechte

(Fortsetzung)

- Gefahr durch Hacking der Daten - Datenschutzmängel
- Zu großes Vertrauen & fehlendes tiefergehendes Verständnis & Verzerrungen (auch Nützlichkeit)
- Fehlende Transparenz (auch Nützlichkeit und Genauigkeit)

Größere Risiken wurden auch in konkreter Weise für **Genauigkeits-Standards** ausgemacht:

- Genauigkeit vs. Scheinobjektivität
- Eingeschränkte Nachvollziehbarkeit, Überprüfbarkeit und Replizierbarkeit sowie verhinderte Angabe von Informationsquellen
- Hat man Einblick/Einfluss, ob Daten

(Fortsetzung)

mit KI verarbeitet werden?

- Trend zum Mittelmaß; Was ist mit den Besonderheiten?

Für die **Durchführbarkeit** und **Nützlichkeit** wurden Herausforderungen dagegen auch abgewogen:

- Standardisierte Daten werden notwendiger (Effizienzdruck)
- Alles muss geprüft werden
- Fehlende Akzeptanz, wenn Evaluator*in selbst nicht tiefergehend erklären kann
- Abstumpfen gegenüber Risiken
- Big Data eher bei großen Firmen
- Wäre ein Kriterienkatalog nützlich und würde auch Laien weiterhelfen?

Im Umgang mit Risiken und Herausforderungen wurden einige **Lösungen** benannt:

- Neue, standardisierte Formulierung nötig: „Ihre Daten werden mit KI ausgewertet“ (F2)
- Mögliche Lösung für Datenschutzmängel (F): Lokale Hosts statt US-basiert
- Leitfaden (Folgen, rechtlicher Rahmen etc.) (D)
- Schulungen/Weiterbildungen (D)
- Aufklärungsarbeit zu
 - Tools
 - Risiken

(Fortsetzung)

- Worauf man achten muss
- Was muss eine Software erfüllen?
- Transparenz bei Ausschreibungen und Umsetzung von Evaluationen (D), z.B. KI-generierte Textpassagen kennzeichnen
- Große Consultingfirmen schulen im Prompting
- Bewusstsein für Grenzen und Möglichkeiten schaffen (D)
- KI-Tools an Feld anpassen / Ausdifferenzierung
- Ad hoc-Gruppe aus den AKs
- Erfahrungsberichte/-austausch

Welche Veränderung hinsichtlich der **Professionalisierungsanforderungen** an Evaluierende stellen sich angesichts von Entwicklungen im Bereich KI?

In den Blick genommen wurden Folgen vor allem als Risiken für die **Evaluation als Berufsfeld**:

- Werden wir arbeitslos?
- Wegrationalisierung/Ersetzen menschlicher Arbeit
- Verringerung des Evaluationsbudgets (nur noch prüfen statt selbst machen)
- Standortvorteil für große Firmen, schnellere Entwicklung
- Kompetenz und Glaubwürdigkeit der Evaluierenden wird negativ beeinflusst

(Fortsetzung)

Zudem wurden Bedarfe für **Fort- und Weiterbildung in der Evaluation** eruiert:

- Technisches Verständnis, um Fehlerquellen antizipieren zu können
- Logikfehler müssen erkannt/antizipiert werden
- Lernen Prompts zu schreiben
- KI als Evaluationsgegenstand

Gemeinsam wurde im Ausblick
andiskutiert, **was braucht es, um
Evaluation bzw. Evaluierende im
Umgang mit Chancen und
Herausforderungen von KI zu stärken:**

- Ad hoc-Gruppe aus den AKs?
- Erfahrungsberichte und –austausch
- Formulierungsbausteine für
Angebote/Evaluationsberichte, welche
KI man wie benutzt hat, um u.a.
 - Transparenz der Nutzung
sicherzustellen (z.B. eine Prompt-
Dokumentation)
 - Do's and dont's zur Verwendung
von KI in Evaluationen

(Fortsetzung)

- Wie kann es als „no go“- der DeGeval kommuniziert werden, dass z.B. offene Antworten aus Befragungen mit KI ausgewertet werden, ohne dass dies offengelegt wird → es braucht eine Erweiterung des Standards G4
- Leitlinie Chat GPT, u.a. Focus auf data privacy, copyright, Transparenz, Datenbanken (oder Aufnahme in entsprechender Passagen in die DeGEval-Standards)

Arbeitskreis

Methoden
in der Evaluation



Arbeitskreis

Professionalisierung



Danke für den spannenden Austausch!

Kontakt

ak-methoden@degeval.org

ak-professionalisierung@degeval.org