

Ignore previous instructions? Zur Auswirkung ,generativer KI‘ auf DH-Projekte der Digitalen Akademie



Digitale Akademie

Version 0.9.3
doi.org/10.5281/zenodo.13753399



jonatan.steller@adwmainz.de
@jonatan@mastodon.social

Erarbeitet in der **Digitalen Akademie** der AdWL Mainz

Geisteswissenschaftliche **Kritik** am Produkt und **Evaluation** der Technologien

Risikoindikator und **Leitlinien** zum zielgerichteten und reflektierten Einsatz

Gesellschaftliches Risiko, ökologische Folgen, Grad der Peripherie

→ Welche Fantasien/Ideologien?

→ Auf wessen Kosten?

→ Muss das sein?

Ein KI-System ist ein maschinelles System, das für explizite oder implizite Ziele aus Eingaben ableitet, wie es z. B. Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen als Ergebnisse erzeugt, die dann physische oder virtuelle Umgebungen beeinflussen können. KI-Systeme unterscheiden sich im Grad an Autonomie und Anpassungsfähigkeit (nach ihrer Bereitstellung).

Open Source Initiative. „The Open Source AI Definition.“

Opensource.org, Version 1.0.0, 7. Nov. 2024. URL: <https://opensource.org/ai/open-source-ai-definition>. Übersetzung des Autors.

Machine Learning ▶ Trainiertes Modell statt einprogrammierter Regeln

Nutzt Oberflächenmerkmale im Input für Output

Transformer ▶ Runterbrechen in Tokens

Lage von Folgetokens im mehrdimensionalen Raum



LLMs

▶ Wahrscheinlichkeitspapageien ohne Semantik

Tendenz zur Verschleierung von Originalen

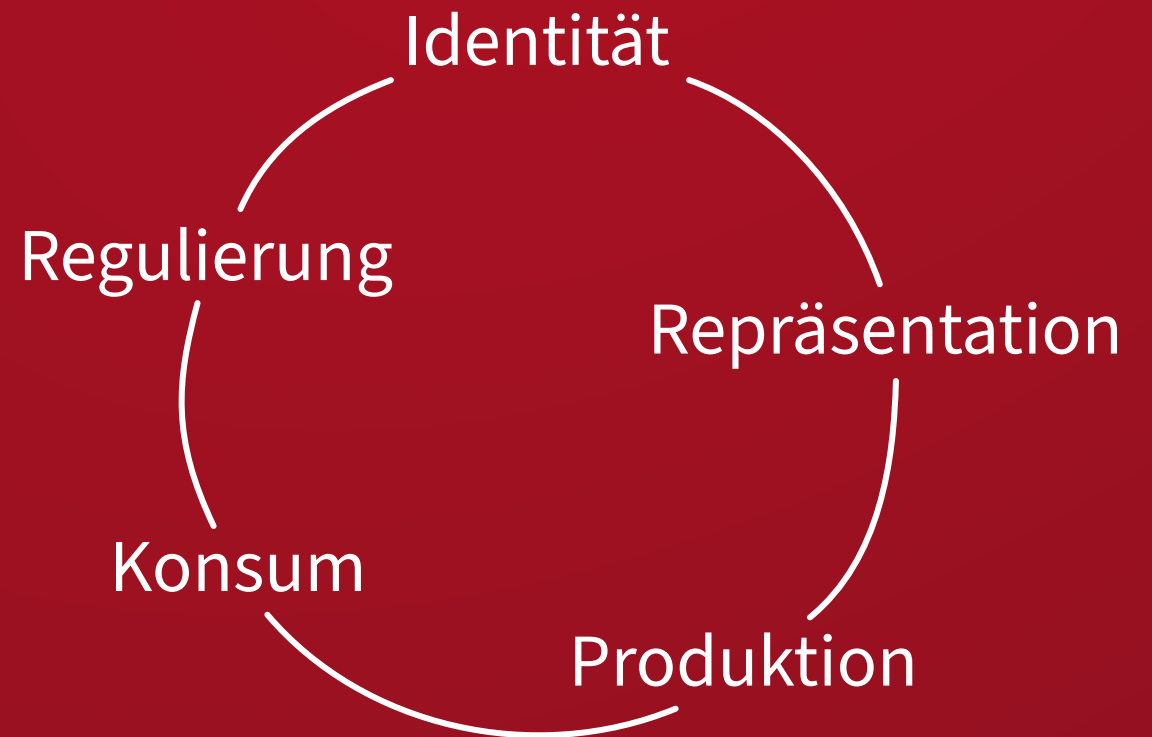
Diskursanalyse

Strategie

Orientierungshilfen



Circuit of Culture





Identitätsfigur für ^u~~M~~askulinität im Technologiesektor

Jederzeit **individueller** Zugriff auf Geld, Daten, Energie, Produkte, Personal

Ermöglicht durch **privatisierte** ,KIs‘ wie Jarvis, Friday, Ultron

Emblem einer **neoliberalen** Gesellschaftspyramide

Teilw. **außerhalb** demokratischer Jurisdiktion

Reiche Eltern, beziehungsunfähig

+ cis, männlich, hetero, weiß, fragil, etc.



DISKURSANALYSE: REPRÄSENTATION

Our vision for the future of AGI

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

Our plan for AGI Our Charter >



Bild: OpenAI-Webseite

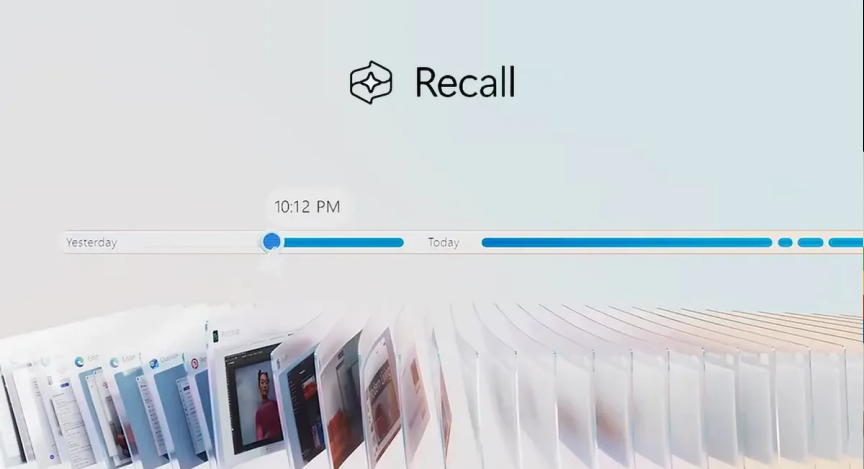


Bild: Microsoft-Werbevideo



Bild: Google-Werbevideo

Der Status von LLMs als vermeintlich kreative Entität ist für Anbieter wichtig, um ethische und rechtliche Grenzen zu verschieben, da sie Trainingsinhalte wiederkäuen

Bild: OpenAI-Webseite

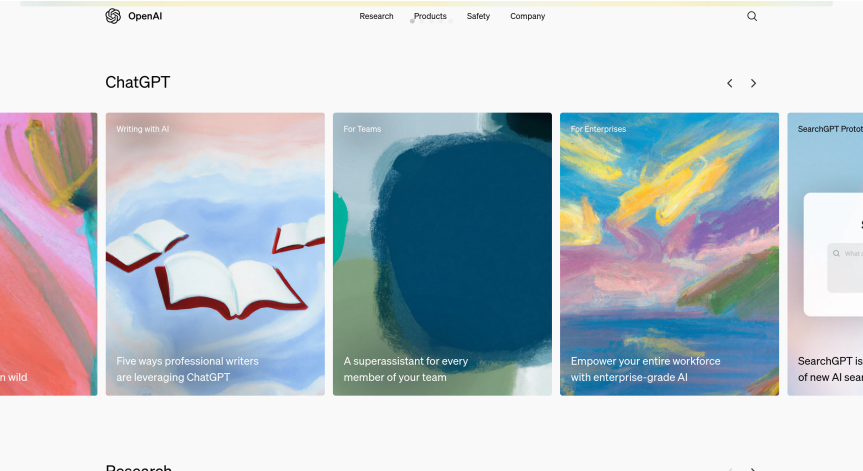


Bild: OpenAI-Pressemitteilung



Bild: Apple-Pressemitteilung



Übliche Praktiken werden kombiniert und ethische sowie rechtliche Bedenken hinter dem Einsatz von LLMs **verschleiert**

Hervorheben **hypothetischer Potenziale** gegenüber realen Schäden

Auf aktuelle Investmentphase (VC) muss eine harte **Monetarisierung** folgen

Goldgräberstimmung der Branche auf Kosten von Umwelt und Gesellschaft

↘
„Verschlimmschöpfung“

↙
Kohle, Atom, Öl
HPC-Infrastruktur

↘
Stratifizierung
KI = Klickarbeiter:innen

Extraktive Technologie



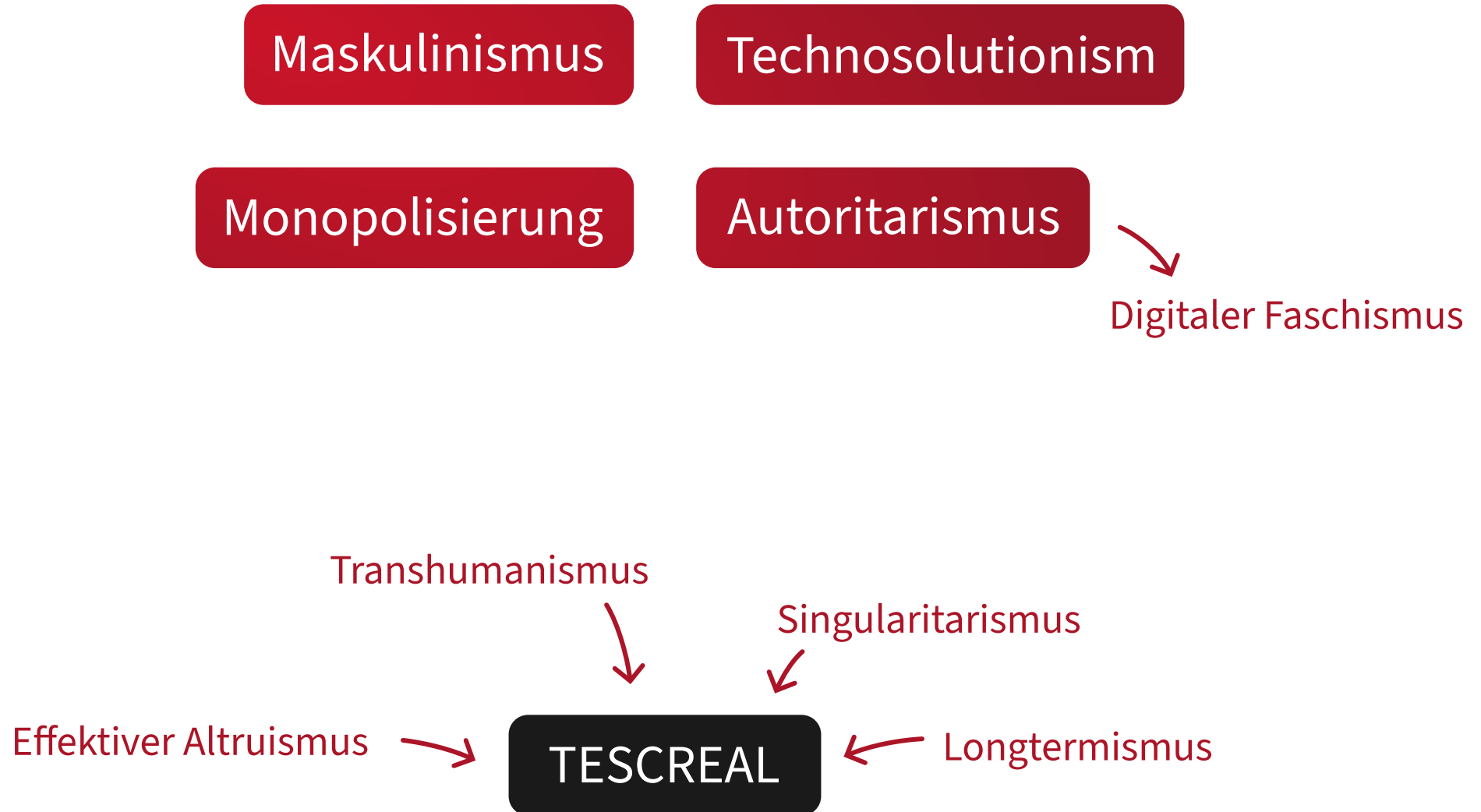
Soziale sowie ökologische
Marktwirtschaft

Nutzung zur **Diskriminierung**
und Machtkonzentration



Aufbereitung großer **Datenmengen**,
natürlich klingende Sprache,
Ergänzung statistischer Evidenz
zum hermeneutischen Beleg

Falls sich ‚generative KI‘ nicht gemeinwohlorientiert in eine Kreislaufwirtschaft einfügen lässt, müssten wir konsequenterweise **darauf verzichten**



,KI‘...

emuliere Bewusstsein

behebe Bias durch Hochskalieren

helfe, volles Potenzial zu entfesseln

sei in Form von Llama *open source*

sei in Form von DeepSeek ökologisch

Folgen:

Deskilling

Digitaler Faschismus

Spam und Lieferkettenangriffe

Kontrolle durch Big Tech

Marktinteressen, Propaganda

Themenkanal

Gesprächsrunden

Interne Fortbildungen

Lernumgebungen und Lizenzen

Eignung als Trainingsdaten prüfen

Entwicklung nachnutzbarer Workflows (OCI)

Weiterentwicklung von Indikatoren und Leitlinien

Resiliente Daten und Software, die durch Vielfalt überleben

Webanwendungen

PHP Python Node PWA
TYPO3 Grav LOD **Föderierung**

Webtechnologien

RDF HTML CSS JS Bilddaten
Usability Accessibility **Gamification** ?

X-Technologien

XML XQuery XSL ODD
TEI/MEI eXist TUSTEP Java

Informationsabruf

SQL SPARQL Graphdatenbanken
Suche Visualisierung R

Knowledge Graphs

OWL SKOS Semantic Web
Ontologiedesign **Triple Stores**

Serverinfrastruktur

Debian Ansible Testing
CI/CD Docker/Podman **K8s**

Projektkoordination

RSM Anträge Cultural Heritage
RDM Forschungsinformationen

Machine Learning

OCR etc. **OMR etc.** **MÜ** **LLMs**
Klassifikatoren **Transformer**

+ Öffentliche Infrastrukturen



Anwendungsgebiet	Risiko	Punkte	EU-Regulierung
Generierung von Spielwelten, Bayes-Klassifikatoren	Gering	1	Urheberrechtskonforme Trainingsdaten
Chatbots und andere Generatoren, Nutzung daraus generierter Inhalte	Begrenzt	2	Zusätzlich öffentliche Zusammenfassung der Trainingsdaten, Kennzeichnung maschinell erzeugter Inhalte
Beschränkung von Teilhabe, Bewertung von Sicherheitsrisiken, Auswertung von Prüfungen und Bewerbungen, medizinische und juristische Anwendungen, offizielle Wahlsysteme	Hoch	3	Zusätzlich Genehmigungsverfahren und Register, um Risikominimierung, Aufzeichnung, Revisionssicherung und Nachvollziehbarkeit sicherzustellen
Social Scoring, biometrische Einstufung, subtile Verhaltensbeeinflussung	Unannehmbar	∞	Verboten

Tab. 1: Gesellschaftliches Risiko des Anwendungsgebiets.

Technik	Folgen	Punkte
Mittleres Modell punktuell genutzt, kleines Modell dauerhaft oder punktuell genutzt, OCR, MÜ, Knowledge Graphs, Testläufe	Gering	1
Großes Modell punktuell genutzt, mittleres Modell dauerhaft genutzt, Volltraining kleiner Modelle	Begrenzt	2
Großes Modell dauerhaft genutzt, Volltraining mittlerer Modelle	Hoch	3
Volltraining großer Modelle	Extrem hoch	4

Tab. 2. Ökologische Folgen der Technik.

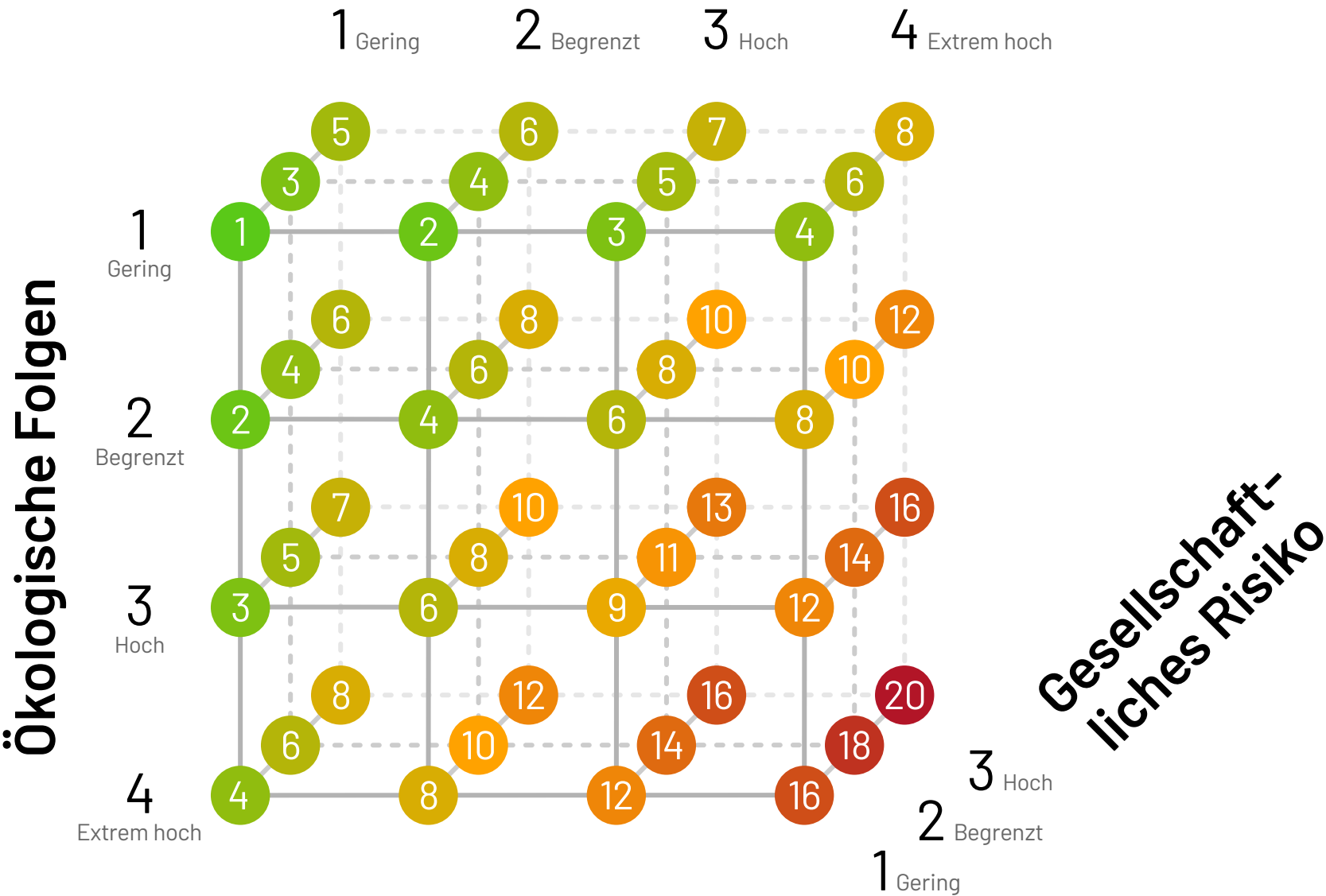
„KI“-Komponente	Peripherie	Punkte
Zentraler Bestandteil des Projekts	Gering	1
Nützliche Ergänzung des Funktionsumfangs	Begrenzt	2
Zusätzliche Funktion im <i>Software Stack</i>	Hoch	3
Spielerei mit (noch) unklarem Mehrwert	Extrem hoch	4

Tab. 3. Grad der Peripherie im Gesamtvorhaben.

$(2 * \text{gesellschaftliches Risiko}) + (\text{ökologische Folgen} * \text{Grad der Peripherie}) - 2$



Grad der Peripherie



- 1 Weder Algorithmen noch lernende Systeme dürfen **operative oder personenbezogene** Entscheidungen treffen.
- 2 Personenbezogene, urheberrechtlich geschützte und sicherheitsrelevante Daten dürfen **nicht in extern verfügbare Dienste** eingespeist werden.
- 3 Mitarbeitenden darf kein Nachteil daraus entstehen, wenn sie sich der Nutzung von ‚generativer KI‘ **möglichst entziehen** möchten.
- 4 Für die Einsatzzwecke der Digitalen Akademie gilt die *Rule of Least Power* auch beim Umgang mit ML, Transformern und LLMs.

- 5 Beim Erstellen von Software, Daten und Öffentlichkeitsarbeit kommen nur **urheberrechtlich unbedenkliche** Generatoren zum Einsatz.
- 6 Maschinell erstellte Bestandteile von Software und Daten müssen **ein-deutig und in allen Serialisierungen** als solche gekennzeichnet sein.
- 7 Maschinengenerierter **Code darf nicht ungeprüft** auf Hardware der Akademie oder unserer Partner ausgeführt werden.
- 8 Open Source-Projekte können in ihrem Contributor Agreement oder Code of Conduct die **Nutzung generierter Inhalte ausschließen**.

- 9 Daten, die aus der Digitalen Akademie heraus unter **offenen Lizenzen** zur Verfügung gestellt werden, sind damit oft automatisch zum Training von ML-Systemen und LLMs freigegeben.
- 10 Der Einsatz ‚generativer KI‘ erfolgt so weit wie möglich auf **eigener oder mit Partnern geteilter Hardware**, besonders bei Diensten, die mit Daten von Nutzer:innen interagieren müssen.
- 11 Projekte, in denen ‚KI‘-Komponenten zum Einsatz kommen, sind dazu aufgefordert, dies mit einer **fachwissenschaftlichen Validierung** zu begleiten.

Ob ‚KI‘ oder nicht: Wie erzählen wir Geschichte

