- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

Die optimale generative KI

Johannes Hötter
Christian Warmuth

Ethische Richtlinien für Künstliche Intelligenz

1. Vorrang menschlichen Handelns und menschlicher Aufsicht
2. Technische Robustheit und Sicherheit
3. Schutz der Privatsphäre und Datenqualitätsmanagement
4. Transparenz und Erklärbarkeit
5. Vielfalt, Nichtdiskriminierung und Fairness
6. Gesellschaftliches und ökologisches Wohlergehen
7. Rechenschaftspflicht

Ethische Leitlinien für Künstliche Intelligenz des BMWK

Ergänzt um folgende Punkte

- Daten bleiben in der Hand der NutzerInnen
- Kostet wenig, schadet der Umwelt nicht
- Ist Open Source und nachvollziehbar
- Denkt sich keine Fakten aus
- Ist nicht manipulierbar
- Es besteht kein Monopol
- Diskriminiert nicht
- Ist sicher und erklärbar

Alignment in der KI: Inner vs Outer Alignment

Alignment in der KI: KI Alignment Forschung zielt darauf ab, KI-Systeme auf vom Menschen angestrebte Ziele zu lenken.

Outer Alignment Problem: Worauf sollten wir unser Modell ausrichten? Spezifizierung einer „Reward"-Funktion, die menschliche Präferenzen für die KI spezifiziert. Problem - schwer zu spezifizieren und Einigung schwer

Inner Alignment Problem: Agiert dieses Modell wirklich nach den spezifizierten Zielen? Wie stellen wir sicher, dass unser Modell, welches auf die Reward Funktion optimiert ist, wirklich unsere von Menschen spezifizierten Zielen erfüllt.

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Überblick über die aktuellen Entwicklungen

Die aktuellen Entwicklungen lassen sich in mehreren Stichpunkten zusammenfassen:

- “Fail often and fast”.
- Viele kurze Iterationen
- Entwicklungen nicht hinter “verschlossenen Türen”
- Entwicklungen werden schnell auf eine breiten Masse “ausgerollt”

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

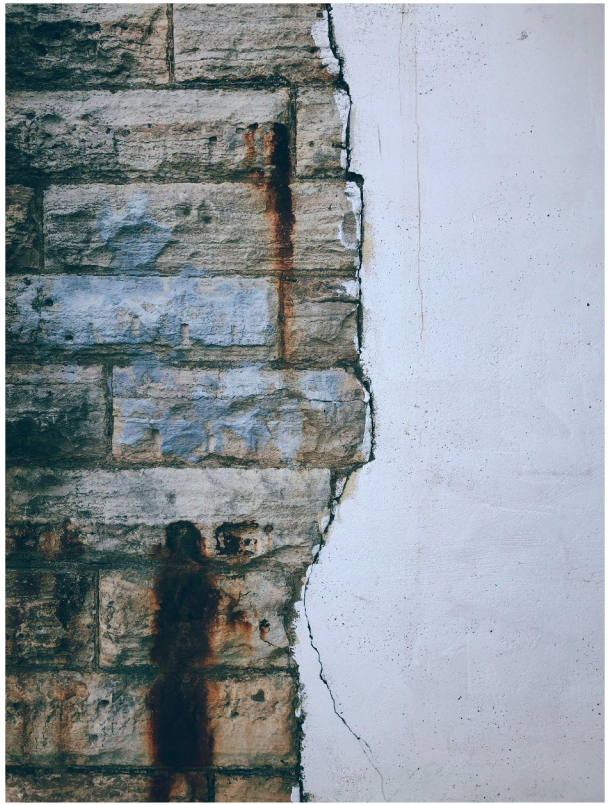
KI Service Zentrum
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Wunsch/Zukunft vs Realität (Stand jetzt)

Vergleich von Wunsch/Zukunft und aktueller Realität in zentralen Themen:

- Copyright
- Open Source
- Halluzinationen
- Jailbreaks
- Bias/Diskriminierung
- Monopolstellung und die Rolle Europas
- KI Sicherheit und Erklärbarkeit
- Nachhaltigkeit und Trainingskosten

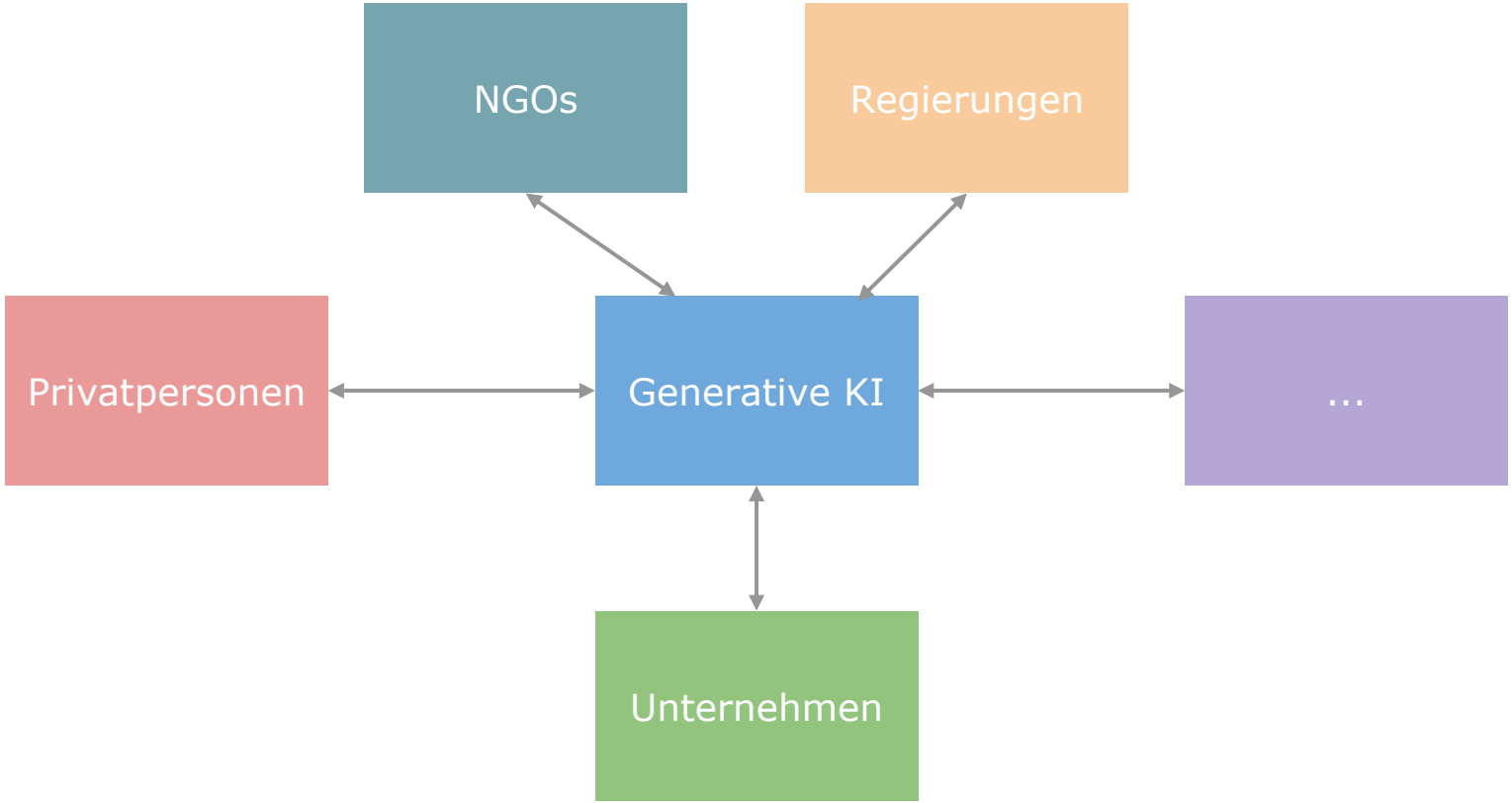


HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum HPI
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Spannungsfeld und unterschiedliche Akteure



Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum
by Hasso-Plattner-Institut
HPI

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Die optimale (generative) KI

Johannes Hötter,
Christian Warmuth

openHPI

- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

Die optimale generative KI

Johannes Hötter
Christian Warmuth