

Let's Shape

the AI Future

Together

KI Ethik in der Verwaltung von Forschungseinrichtungen

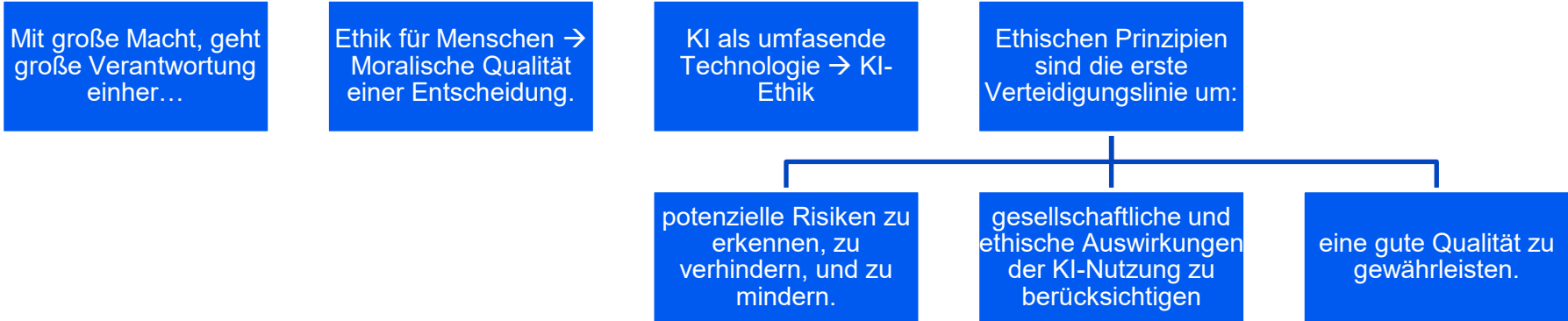
Grundlagen von Trustworthy AI

Rodrigo Conde Jiménez

AIT

25.06.2026

KI-Ethik: Wozu geht es?



In der Forschungseinrichtungen...



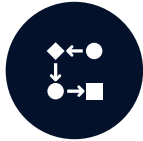
Hervorragende Qualität
der wissenschaftlichen
Tätigkeiten



Achtung und
Einbeziehung
unterschiedlichen
Perspektiven



Wissenschaftliche
Wahrheitsfindung als
unverzichtbares Prinzip



Transparente Prozesse
und gebührende
Sorgfaltspflicht



Gute
Forschungsgovernance
→ Gute KI-Governance

Ein paar berühmte Beispiele

- Nature → KI-Skandal (erfundene Zitate)
- Paper Mills
- Gefälschte Peer Reviews
- Manipulation visueller Daten
- KI-Betrugsfall in Indonesien

nature

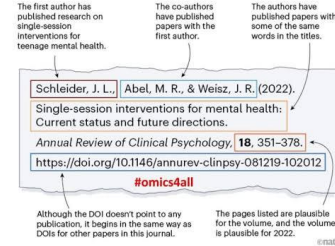
Hallucinated citations are polluting the scientific literature. What can be done?

Tens of thousands of publications from 2025 might include invalid references generated by AI, a Nature analysis suggests.

HOW FAKES CAN LOOK REAL

This citation, generated by AI as part of a 2025 study, looks plausible even though it points to research that doesn't exist.

#omics4all



Ein paar berüchtigte Beispiele

- Nature → KI-Skandal (erfundene Zitate)
- Paper Mills
- Gefälschte Peer Reviews
- Manipulation visueller Daten
- KI-Betrugsfall in Indonesien

nature

Hallucinated citations are polluting the scientific literature. What can be done?

Tens of thousands of publications from 2025 might include invalid references generated by AI, a Nature analysis suggests.

HOW FAKES CAN LOOK REAL

This citation, generated by AI as part of a 2025 study, looks plausible even though it points to research that doesn't exist.

#omics4all

The first author has published research on single-session interventions for teenage mental health.

The co-authors have published papers with the first author.

The authors have published papers with some of the same words in the titles.

Schleider, J. L., Abel, M. R., & Weisz, J. R. (2022).

Single-session interventions for mental health: Current status and future directions.

Annual Review of Clinical Psychology, 18, 351-378.

<https://doi.org/10.1146/annurev-clinpsy-081219-102012>

#omics4all

Although the DOI doesn't point to any publication, it begins in the same way as DOIs for other papers in this journal.

The pages listed are plausible for the volume, and the volume is plausible for 2022.

Artificial intelligence Tech / Tech Trends

AI-powered fraud: Chinese paper mills are mass-producing fake academic research

Paper mills that sell authorship or fabricate entire papers are a staple of China's competitive academic landscape

Reading Time: 2 minutes



Ein paar berühmte Beispiele

- Nature → KI-Skandal (erfundene Zitate)
- Paper Mills
- Gefälschte Peer Reviews
- Manipulation visueller Daten
- KI-Betrugsfall in Indonesien

nature

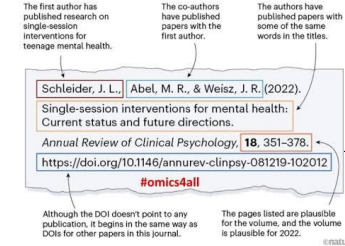
Hallucinated citations are polluting the scientific literature. What can be done?

Tens of thousands of publications from 2025 might include invalid references generated by AI, a *Nature* analysis suggests.

HOW FAKES CAN LOOK REAL

This citation, generated by AI as part of a 2025 study, looks plausible even though it points to research that doesn't exist.

#omics4all



Artificial intelligence Tech / Tech Trends

AI-powered fraud: Chinese paper mills are mass-producing fake academic research

Paper mills that sell authorship or fabricate entire papers are a staple of China's competitive academic landscape

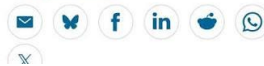
Reading Time: 2 minutes



Major AI conference flooded with peer reviews written fully by AI

Controversy has erupted after 21% of manuscript reviews for an international AI conference were found to be generated by artificial intelligence.

By Miryam Naddaf



Ein paar berüchtigte Beispiele

- Nature → KI-Skandal (erfundene Zitate)
- Paper Mills
- Gefälschte Peer Reviews
- Manipulation visueller Daten
- KI-Betrugsfall in Indonesien

nature

Hallucinated citations are polluting the scientific literature. What can be done?

Tens of thousands of publications from 2025 might include invalid references generated by AI, a Nature analysis suggests.

HOW FAKES CAN LOOK REAL

This citation, generated by AI as part of a 2025 study, looks plausible even though it points to research that doesn't exist.

#omics4all

The first author has published research on single-session interventions for teenage mental health.

The co-authors have published papers with the first author.

The authors have published papers with some of the same words in the titles.

Schleider, J. L., Abel, M. R., & Weisz, J. R. (2022).
Single-session interventions for mental health:
Current status and future directions.
Annual Review of Clinical Psychology, 18, 351-378.
<https://doi.org/10.1146/annurev-clinpsy-081219-102012>

#omics4all

Although the DOI doesn't point to any publication, it begins in the same way as DOIs for other papers in this journal.

The pages listed are plausible for the volume, and the volume is plausible for 2022.

Artificial intelligence Tech / Tech Trends

AI-powered fraud: Chinese paper mills are mass-producing fake academic research

Paper mills that sell authorship or fabricate entire papers are a staple of China's competitive academic landscape

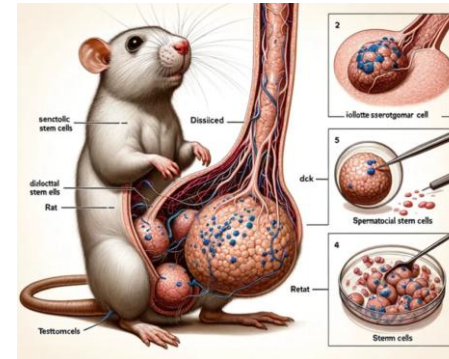
Reading Time: 2 minutes

T

Major AI conference flooded with peer reviews written fully by AI

Controversy has erupted after 21% of manuscript reviews for an international AI conference were found to be generated by artificial intelligence.

By Miryam Naddaf



Ein paar berühmte Beispiele

- Nature → KI-Skandal (erfundene Zitate)
- Paper Mills
- Gefälschte Peer Reviews
- Manipulation visueller Daten
- KI-Betrugsfall in Indonesien

nature

Hallucinated citations are polluting the scientific literature. What can be done?

Tens of thousands of publications from 2025 might include invalid references generated by AI, a Nature analysis suggests.

HOW FAKES CAN LOOK REAL

This citation, generated by AI as part of a 2025 study, looks plausible even though it points to research that doesn't exist.

#omics4all

The first author has published research on single-session interventions for teenage mental health.

The co-authors have published papers with the first author.

The authors have published papers with some of the same words in the titles.

Schleider, J. L., Abel, M. R., & Weisz, J. R. (2022).
Single-session interventions for mental health:
Current status and future directions.
Annual Review of Clinical Psychology, 18, 351-378.
<https://doi.org/10.1146/annurev-clinpsy-081219-102012>

#omics4all

Although the DOI doesn't point to any publication, it begins in the same way as DOIs for other papers in this journal.

The pages listed are plausible for the volume, and the volume is plausible for 2022.

Artificial intelligence Tech / Tech Trends

AI-powered fraud: Chinese paper mills are mass-producing fake academic research

Paper mills that sell authorship or fabricate entire papers are a staple of China's competitive academic landscape

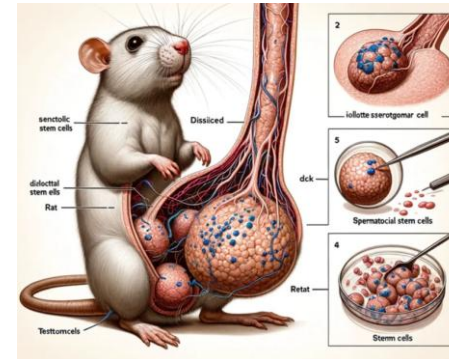
Reading Time: 2 minutes

T

Major AI conference flooded with peer reviews written fully by AI

Controversy has erupted after 21% of manuscript reviews for an international AI conference were found to be generated by artificial intelligence.

By Miryam Naddaf



In der Forschungseinrichtungen...

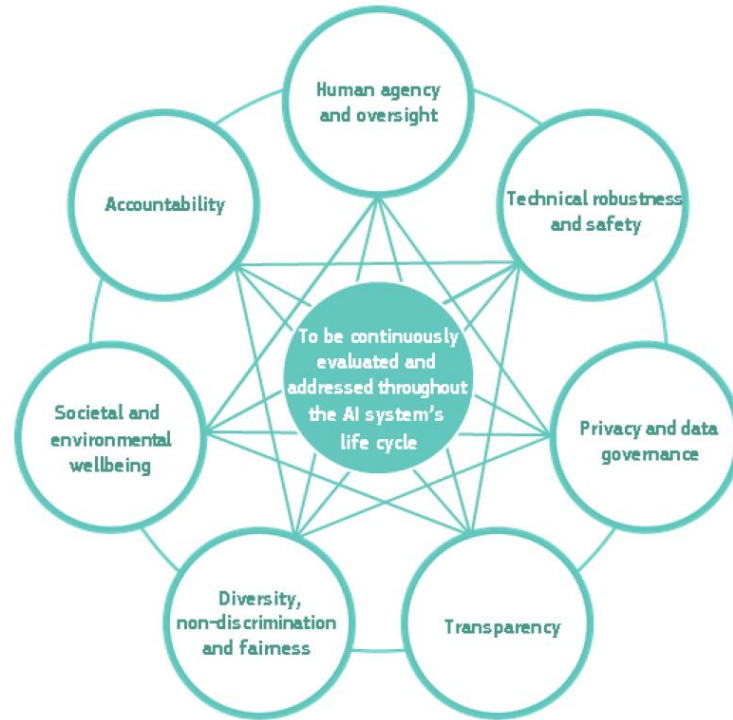
- Wie stellen wir sicher, dass wir eine gute Qualität unserer Arbeit in der „Post-Truth“-Ära gewährleisten?
- Welche Prinzipien sollten in Bezug auf KI bei der Verwaltung einer Forschungseinrichtung berücksichtigt werden?

Im Forschungskontext...

- **Forschungsintegrität:**
Einhaltung wissenschaftlicher Standards durch transparente Methoden, nachvollziehbare Datenverarbeitung und korrekte Autorenschaft sowie Vermeidung von Fälschung, Manipulation und Plagiat.
- **Forschungssicherheit:**
Schutz von Forschungsdaten, digitalen Infrastrukturen und Ergebnissen vor unbefugtem Zugriff, Missbrauch sowie wirtschaftlicher oder politischer Einflussnahme.
- **Epistemische Verantwortung:**
Verantwortung für die Qualität und Begründbarkeit von Wissen unter Berücksichtigung von Unsicherheit, Datenqualität, Modellgrenzen und möglichen Verzerrungen KI-gestützter Forschung.



Im Europa: HLEGs Prinzipien für Trustworthy AI



Im Europa: HLEGs Prinzipien für Trustworthy AI



Menschliche
Aufsicht &
Kontrolle



Technische
Robustheit &
Sicherheit



Datenschutz &
Daten
Governance



Transparenz



Vielfalt,
Nichtdiskriminierung & Fairness



Gesellschaftliches &
Ökologisches
Wohl



Rechenschaftspflicht

Menschliche Aufsicht & Kontrolle



- KI-Anwendungen müssen den Menschen achten und stärken.
- Hinter jeder KI-Anwendung steht eine verantwortliche Person.

Menschliche Aufsicht & Kontrolle



- KI-Anwendungen müssen den Menschen achten und stärken.
- Hinter jeder KI-Anwendung steht eine verantwortliche Person.
- Wichtige Aspekte:
 - Sinnvolle menschliche Kontrolle: HITL-HOTL-HIC
 - Möglichkeit zum Eingreifen und Überstimmen
 - Automation bias, übermäßiges Abhängigkeiten

Menschliche Aufsicht & Kontrolle



- KI-Anwendungen müssen den Menschen achten und stärken.
- Hinter jeder KI-Anwendung steht eine verantwortliche Person.
- Wichtige Aspekte:
 - Sinnvolle menschliche Kontrolle: HITL-HOTL-HIC
 - Möglichkeit zum Eingreifen und Überstimmen
 - Automation bias, übermäßiges Abhängigkeiten
- Beispiel: Bewertung von Förderanträgen

Technische Robustheit & Sicherheit

- KI-Systeme müssen widerstandsfähig, sicher und zuverlässig sein.
- Notfallpläne für den Fall, dass etwas schiefgeht, sind erforderlich.



Technische Robustheit & Sicherheit

- KI-Systeme müssen widerstandsfähig, sicher und zuverlässig sein.
- Notfallpläne für den Fall, dass etwas schiefgeht, sind erforderlich.
- Wichtige Aspekte:
 - Schutz von Angriffe; Resilienz des Systems
 - Genauigkeit, Zuverlässigkeit
 - Access-Control-Policies
 - Beobachtung potenzieller Risiken

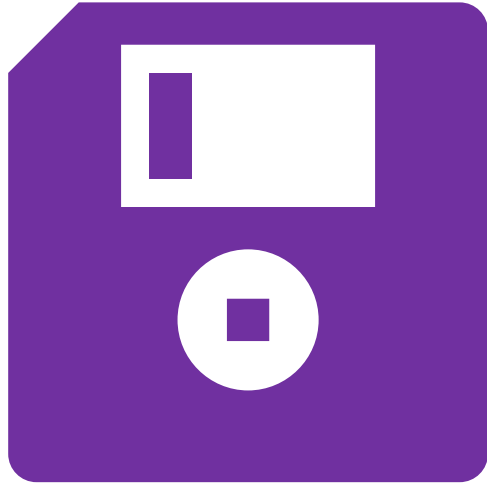


Technische Robustheit & Sicherheit

- KI-Systeme müssen widerstandsfähig, sicher und zuverlässig sein.
- Notfallpläne für den Fall, dass etwas schiefgeht, sind erforderlich.
- Wichtige Aspekte:
 - Schutz von Angriffe; Resilienz des Systems
 - Genauigkeit, Zuverlässigkeit
 - Access-Control-Policies
 - Beobachtung potenzieller Risiken
- **Beispiel:** Verarbeitung von Rechnungen / Verwaltungsdokumenten

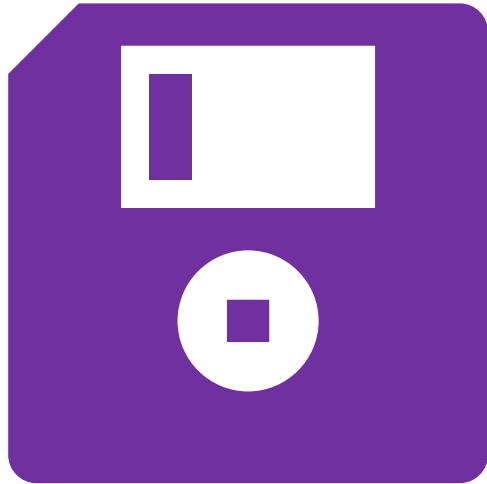


Datenschutz & Daten Governance



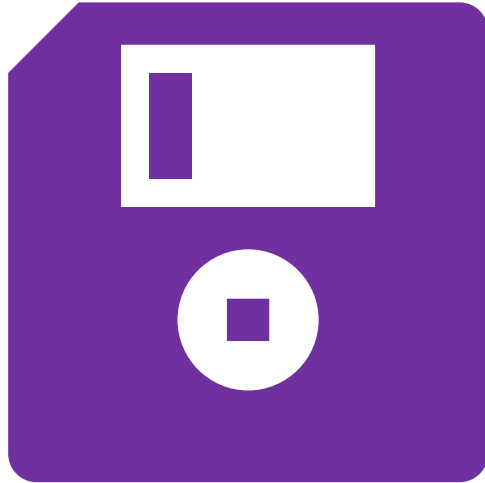
- KI-Tools müssen die individuellen Datenrechte wahren.
- Dazu gehören der Schutz personenbezogener Daten die Sicherung von Trainingsdaten und Datenverwaltung (DGSVO).

Datenschutz & Daten Governance



- KI-Tools müssen die individuellen Datenrechte wahren.
- Dazu gehören der Schutz personenbezogener Daten die Sicherung von Trainingsdaten und Datenverwaltung (DGSVO).
- Wichtige Aspekte:
 - Datenqualität und Integrität
 - Speicherung/Aufbewahrung
 - Datenzugang → Daten-Governance

Datenschutz & Daten Governance



- KI-Tools müssen die individuellen Datenrechte wahren.
- Dazu gehören der Schutz personenbezogener Daten die Sicherung von Trainingsdaten und Datenverwaltung (DGSVO).
- Wichtige Aspekte:
 - Datenqualität und Integrität
 - Speicherung/Aufbewahrung
 - Datenzugang → Daten-Governance
- Beispiel: KI für Personal und Bewerbungsdaten

Transparenz

- Daten, Methoden, System und Zweck sind transparent und nachvollziehbar dokumentiert.



Transparenz

- Daten, Methoden, System und Zweck sind transparent und nachvollziehbar dokumentiert.
- Wichtige Aspekte:
 - Nachvollziehbarkeit (Modell? Daten? Version? Prozess?)
 - Erklärbarkeit (Warum dieses Ergebniss?)
 - Verständlichkeit

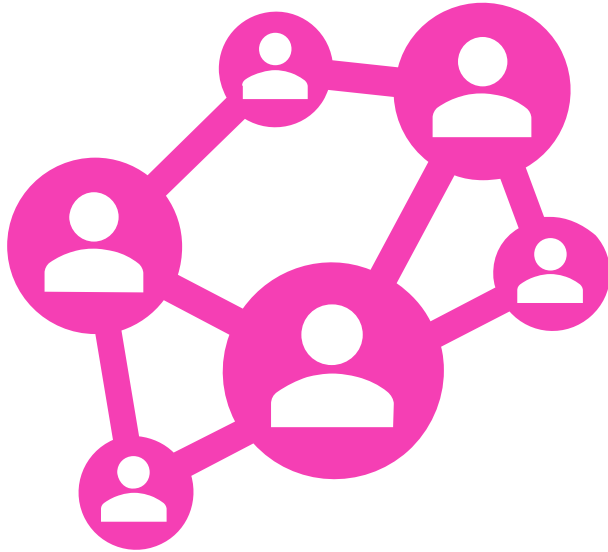


Transparenz

- Daten, Methoden, System und Zweck sind transparent und nachvollziehbar dokumentiert.
- Wichtige Aspekte:
 - Nachvollziehbarkeit (Modell? Daten? Version? Prozess?)
 - Erklärbarkeit (Warum dieses Ergebniss?)
 - Verständlichkeit
- Beispiel: Automatische KI-Bewertung eines Projektantrags

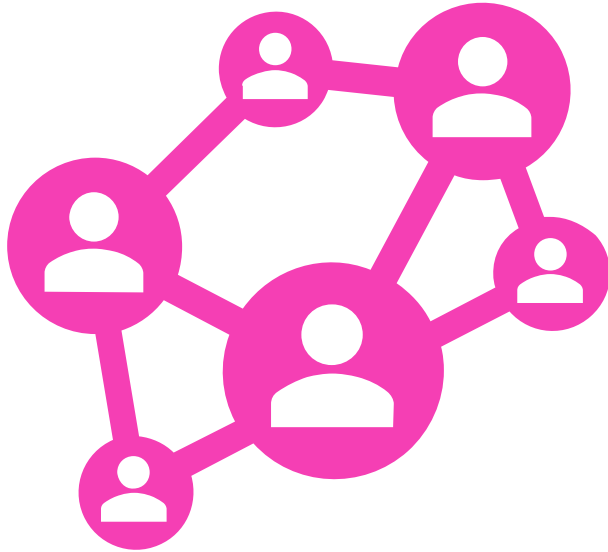


Vielfalt, Nichtdiskriminierung & Fairness



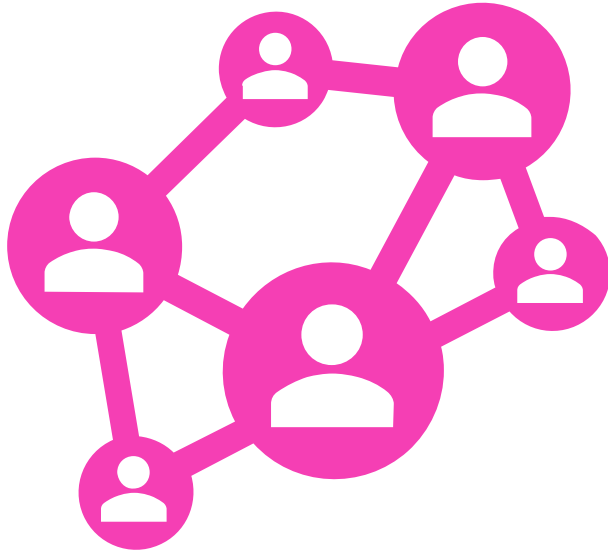
- Systeme müssen Einzelpersonen und Gruppen gerecht behandeln.
- Bei der Entwicklung und Einsetzung sollten Verzerrungen in Datensätzen oder Algorithmen aktiv erkannt, gemessen und verringert werden, um diskriminierende Auswirkungen zu verhindern.

Vielfalt, Nichtdiskriminierung & Fairness



- Systeme müssen Einzelpersonen und Gruppen gerecht behandeln.
- Bei der Entwicklung und Einsetzung sollten Verzerrungen in Datensätzen oder Algorithmen aktiv erkannt, gemessen und verringert werden, um diskriminierende Auswirkungen zu verhindern.
- Wichtige Aspekte:
 - Bias/Verzerrungen (Daten, Algorithmen, Organisation).
 - Repräsentation verschiedenen Gruppen
 - Zugang zur Anwendung/System

Vielfalt, Nichtdiskriminierung & Fairness



- Systeme müssen Einzelpersonen und Gruppen gerecht behandeln.
- Bei der Entwicklung und Einsetzung sollten Verzerrungen in Datensätzen oder Algorithmen aktiv erkannt, gemessen und verringert werden, um diskriminierende Auswirkungen zu verhindern.
- Wichtige Aspekte:
 - Bias/Verzerrungen (Daten, Algorithmen, Organisation).
 - Repräsentation verschiedenen Gruppen
 - Zugang zur Anwendung/System
- Beispiel: Identifikation “vielsprechender” Forschungsprojekte oder Wissenschaftler*innen für interne Förderung

Gessellschaftliches & Ökologisches Wohl

- KI-Systeme müssen allen Menschen – auch zukünftigen Generationen – zugutekommen und andere Lebensformen berücksichtigen.



Gessellschaftliches & Ökologisches Wohl

- KI-Systeme müssen allen Menschen – auch zukünftigen Generationen – zugutekommen und andere Lebensformen berücksichtigen.
- Wichtige Aspekte:
 - Ökologische Nachhaltigkeit (Energieverbrauch, Effizienz)
 - Soziale und gesellschaftliche Auswirkungen



Gessellschaftliches & Ökologisches Wohl

- KI-Systeme müssen allen Menschen – auch zukünftigen Generationen – zugutekommen und andere Lebensformen berücksichtigen.
- Wichtige Aspekte:
 - Ökologische Nachhaltigkeit (Energieverbrauch, Effizienz)
 - Soziale und gessellschaftliche Auswirkungen
- Beispiel: Nutzung generativer KI oder LLMs im Alltag



Rechenschaftspflicht



- Es müssen Mechanismen existieren, um Verantwortung und Rechenschaftspflicht für KI-Systeme und deren Ergebnisse sicherzustellen.

Rechenschaftspflicht



- Es müssen Mechanismen existieren, um Verantwortung und Rechenschaftspflicht für KI-Systeme und deren Ergebnisse sicherzustellen.
- Wichtige Aspekte:
 - Monitoring
 - Gute Aufteilung der Verantwortlichkeiten
 - Risikomanagement
 - Korrektur und Wiedergutmachung

Rechenschaftspflicht



- Es müssen Mechanismen existieren, um Verantwortung und Rechenschaftspflicht für KI-Systeme und deren Ergebnisse sicherzustellen.
- Wichtige Aspekte:
 - Monitoring
 - Gute Aufteilung der Verantwortlichkeiten
 - Risikomanagement
 - Korrektur und Wiedergutmachung
- Beispiel: KI-Chatbot für interne Verwaltungsprozesse.

Ethische prinzipien im Praxis

- Use-Case-Beispiel:
Literaturbeobachtung &
Publikationsmanagement Tool
- Use-Case-Beispiel: Frist- & Daten-
Überwachung Tool

Literaturbeobachtung & Publikationsmanagement Tool

- Vorteile:
 - Schnellerer Überblick über relevante neue Forschung
 - Zeitersparnis bei der Erstellung von Publikationslisten und Berichten
 - Höhere Datenqualität und Konsistenz
- Potenzielle ethische Überlegungen:
- Menschliche Aufsicht:
 - Abhängigkeiten
 - De-Skilling wissenschaftlicher Arbeit
- Vielfalt, ND, Fairness:
 - Publikationen aus bestimmten Wissenskontexten könnten nicht repräsentiert werden.
 - Wer entscheidet was gilt als „relevant“?
- Epistemische Verantwortung:
 - Wissen kann homogener und standardisierter werden
→ allmählicher Verlust der Innovationsbereitschaft



Frist- & Daten-Überwachungs-Tool

- **Vorteile**
 - Wenige verpassten Fristen
 - Automatisierung der Fristlegung
 - Überblick von allen laufenden Fristen
- **Potenzielle ethische Überlegungen:**
- **Menschliche Aufsicht:**
 - Automation Bias
 - Abhängigkeiten vom Tool
- **Daten-Governance:**
 - Welche Daten werden verarbeitet?
 - Sensible Informationen können da stehen
- **Technische Robustheit:**
 - Wie vertrauenswürdig ist die Anwendung?
 - Was passiert, wenn ein Datum falsch eingetragen/analysiert wird?



Verantwortungsvoller Umgang mit KI

Gestaltung in der KI-Entwicklung

- Orientierung am neuesten Stand von Forschung und Technik
- Anwendung anerkannter Standards und Best Practices
- Verantwortungsvolle Datenbasis

Beschaffung und Implementierung

- Due Diligence der Anbieter
- Transparenz und Nachvollziehbarkeit bei der Beschaffung
- Klare Regeln zu zulässigen Anwendungsfällen (AI Policy)
- Vermeidung unnötiger Automatisierung

Zentrale Anforderungen

- Begründeter Bedarf für KI-Einsatz
- Klare Rollen und Verantwortlichkeiten
- Kritische Reflexion (Novelty Bias, Überabhängigkeit)
- Förderung von KI-Kompetenz und reflektierter KI-Nutzung im Forschungsalltag

Best-Practice-Beispiele

- Ethik-Boards: Einrichtung von KI-spezifischen Ethik-Boards oder Erweiterung bestehender Ethik-Boards für KI-Fragen.
- Interdisziplinäre KI-Arbeitsgruppen:
 - Beispiel: AI TaskForce am AIT
- Umsetzung spezifischer KI-Guidelines
 - Beispiel: *“Living guidelines on the responsible use of generative AI in research”*
 - *“WIFO AI Guidelines on the Proper Use of Generative Models of Artificial Intelligence”*



WIFO RESEARCH BRIEFS
9/2025

WIFO AI Guidelines on the Proper
Use of Generative Models
of Artificial Intelligence



AI:AT Consortium

Beneficiaries



Affiliated Entities



Funding



EuroHPC
Joint Undertaking

 Federal Ministry
Innovation, Mobility
and Infrastructure
Republic of Austria



**Funded by
the European Union**



AI Factory Austria is funded by the European High-Performance Computing Joint Undertaking (JU) under Grant Agreement No. 101253078.

Contact



**Rodrigo Conde
Jimenez**

Rodrigo.conde-jimenez@ai-at.eu



info@ai-at.eu



ai-at.eu



[@ai-factory-austria](https://www.linkedin.com/company/ai-factory-austria)

AI:AT AI Factory Austria
Schwarzenbergplatz 2
1010 Wien, Austria