KI-Unterstützung zur Verarbeitung von Text- und Umgebungsdaten in der Gefährdungsbeurteilung

Inhaltsverzeichnis

- Übersicht: Künstliche Intelligenz in der Gefährdungsbeurteilung
- Große Sprachmodelle zur Unterstützung der Gefährdungsbeurteilung

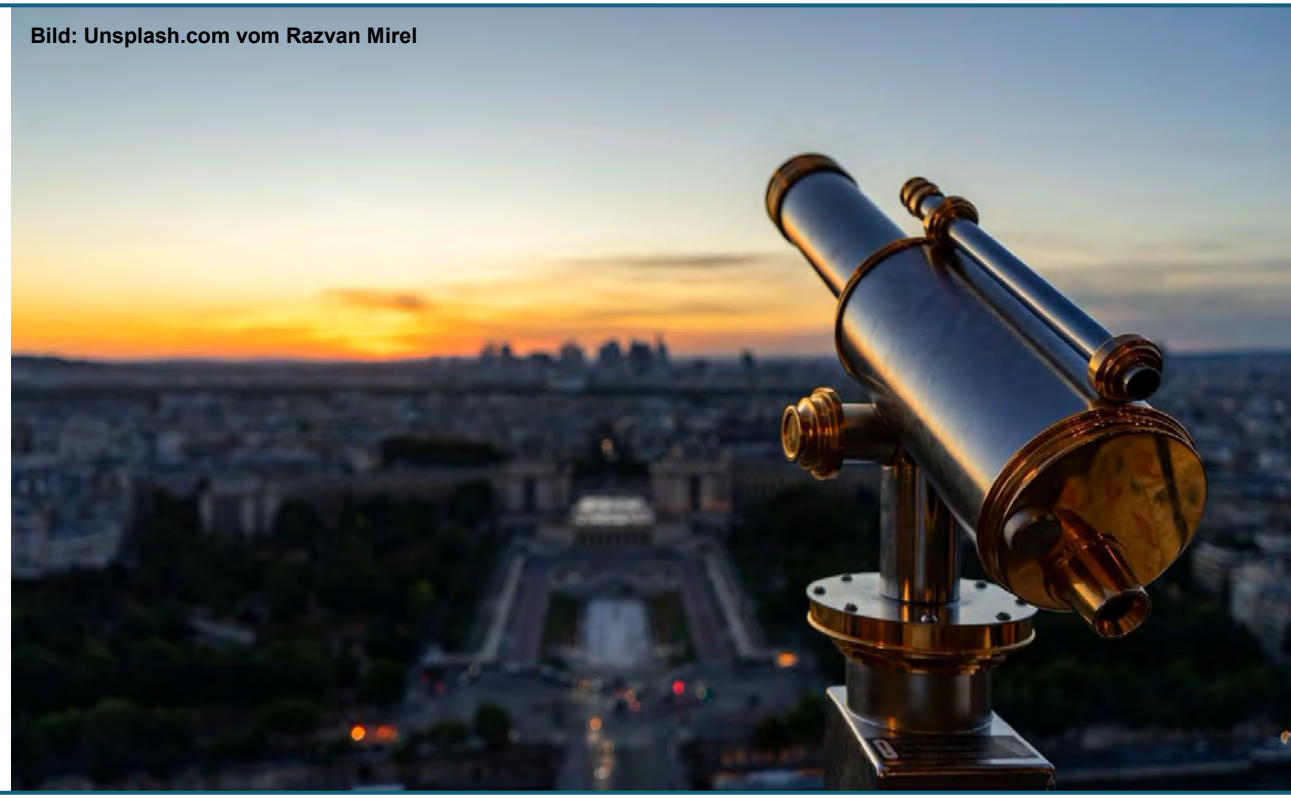


Bild: Unsplash.com vom Razvan Mirel

KI gestützte Gefährdungsbeurteilung an der BAuA

Expositionsabschätzung



Convolutional Neural Networks,
Regressions-Setting

Asbestfaser Detektion

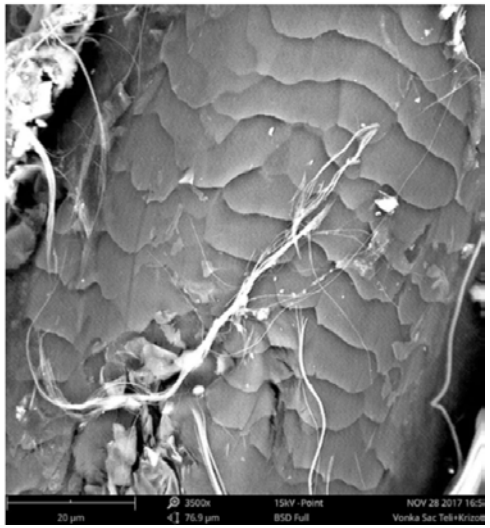
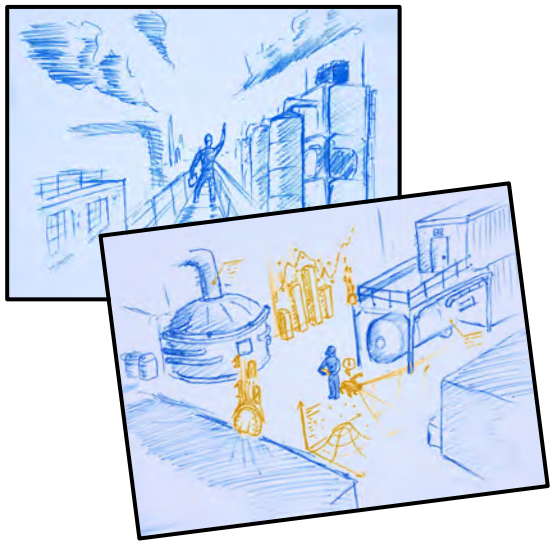


Bild: wikimedia commons von AgtVonka

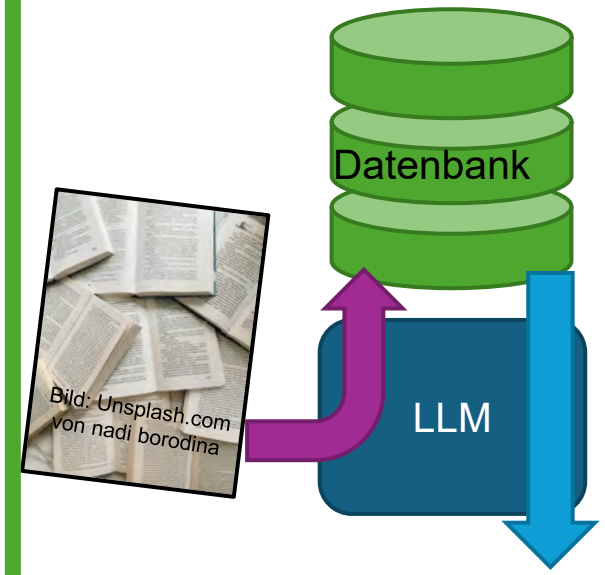
Convolutional Neural Networks,
Klassifikations-Setting

Sensorassistierte GBU



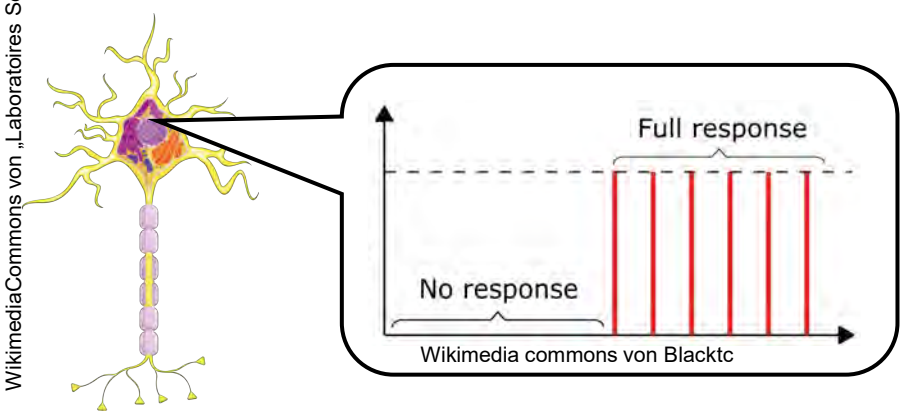
Autoencoders,
Zeitreihen-anomalie-detektion

Chatbots für die GBU



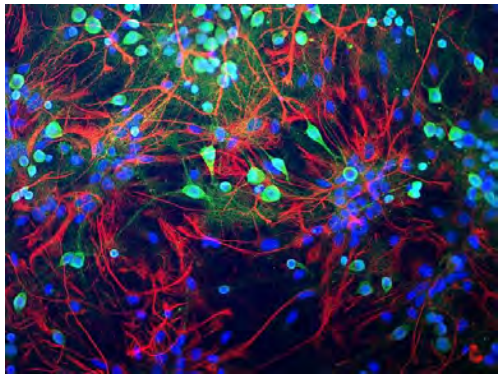
Large Language Models,
Generierung der GBU

Beispiel: Künstliche Neuronale Netze

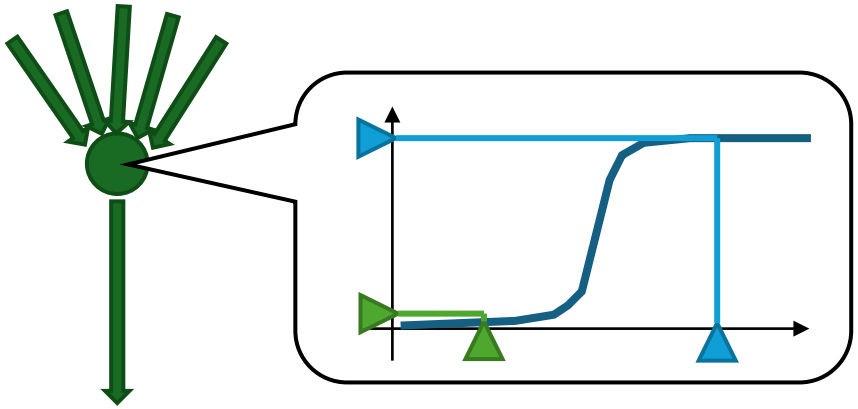


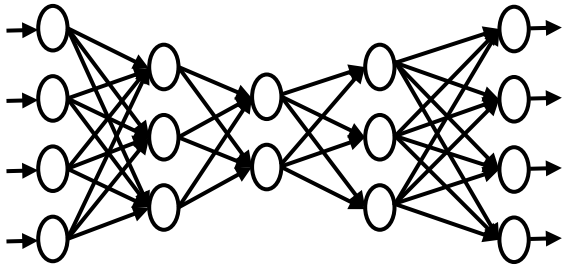
WikimediaCommons von „Laboratoires Servier“

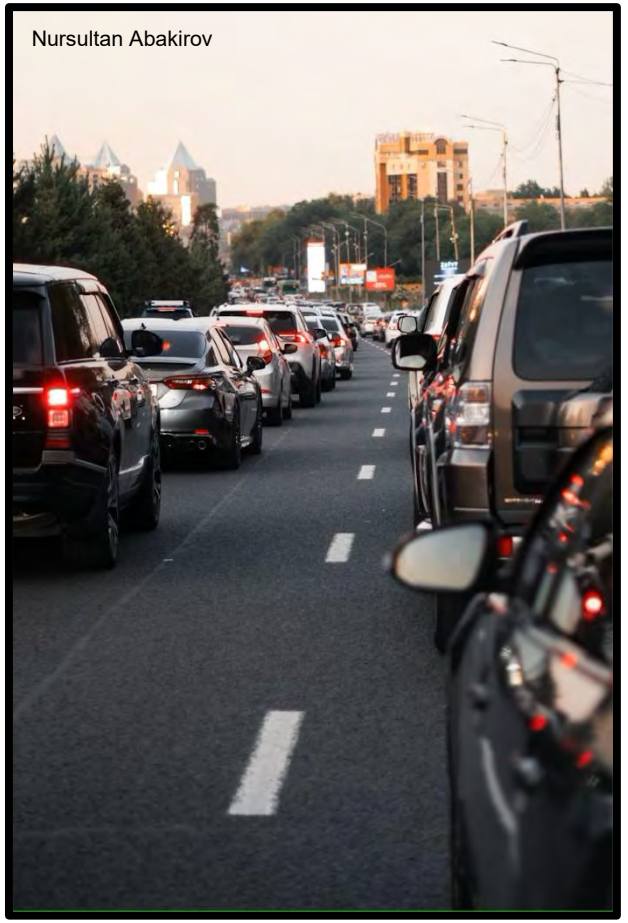
Wikimedia commons von Blacktc



WikimediaCommons von GerryShaw.

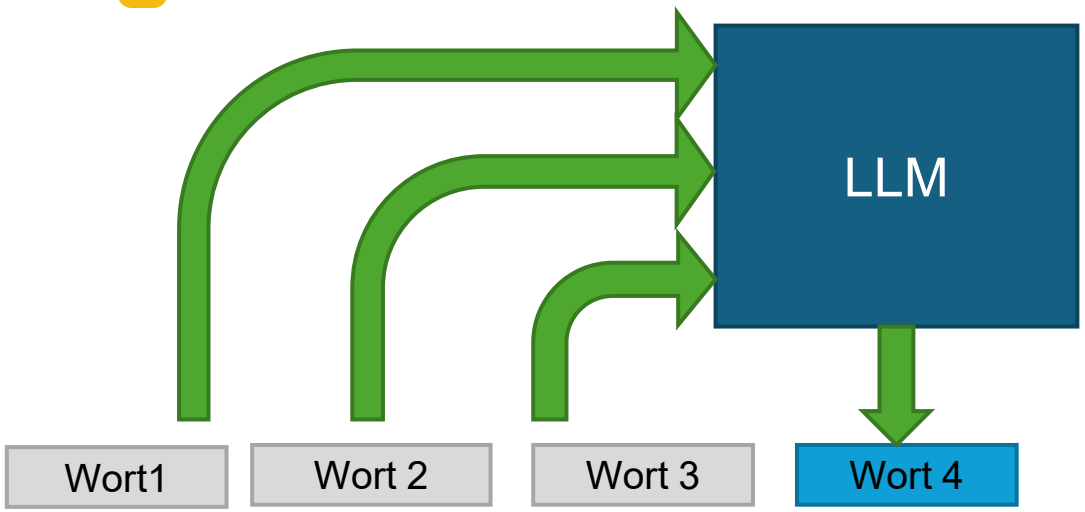






Nursultan Abakirov

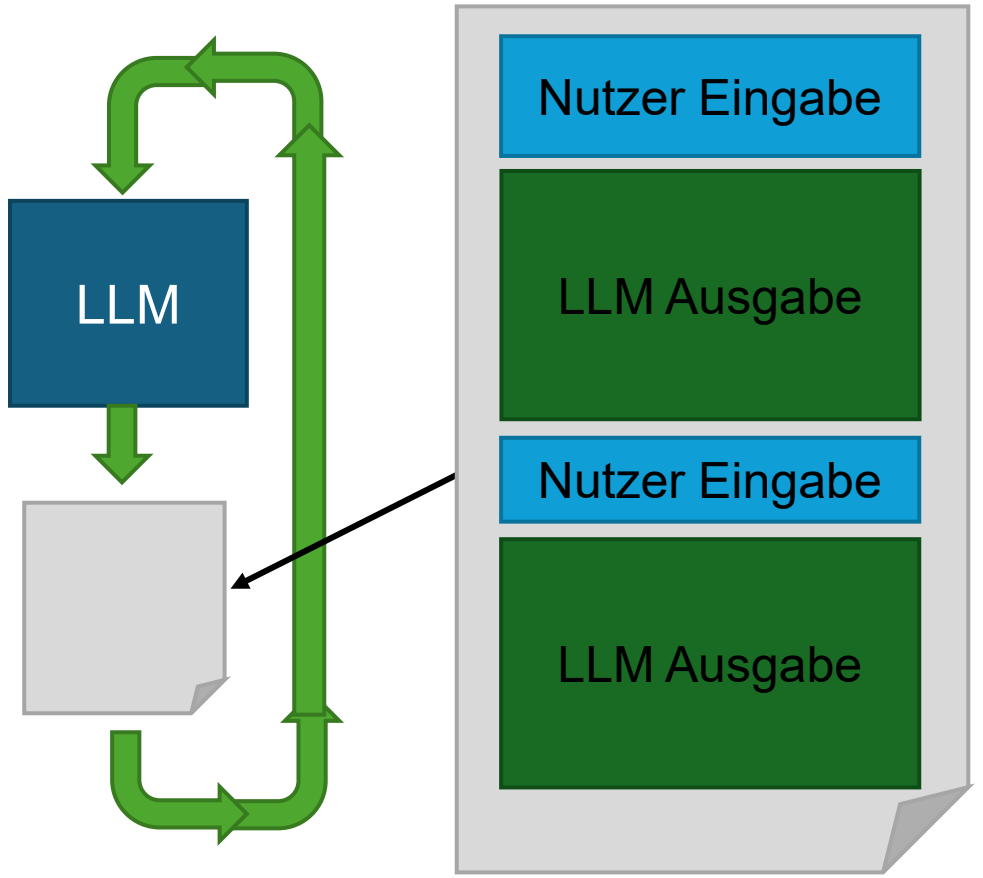
LLM



The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

writings.stephenwolfram.com



Risikobehaftete Eigenschaften von KI



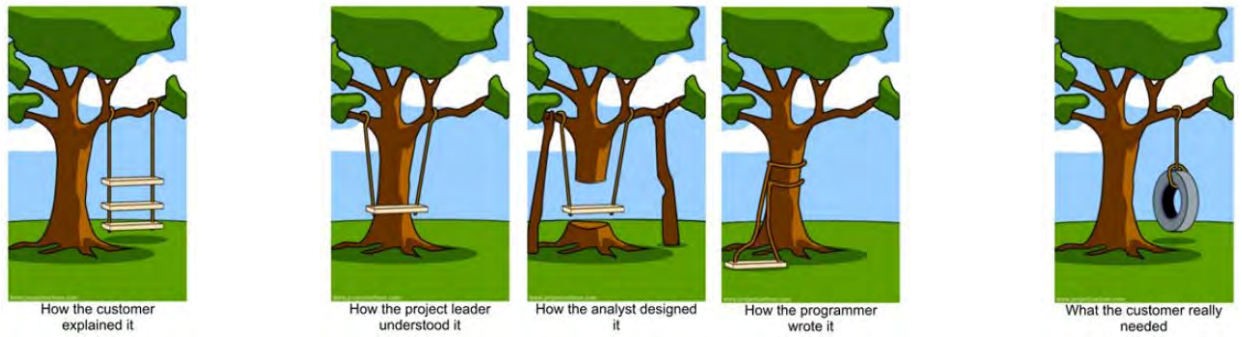
If you can do it without AI, so much the better.

Cassie Kozyrkow

Ex-Chief Decision Scientist, Google

Mutter aller Fehlerquellen: Spezifikationen

- **Spezifikationen:**
  - Ideal: Der unterliegende Wunschzustand
  - Formulierung: Die Ausformulierung dieses Zustandes **Trainingsdaten!**
  - Offenbar: Die Umsetzung des formulierten Zustandes



Große Sprachmodelle zur Unterstützung der Gefährdungsbeurteilung

Gefährdungsbeurteilungen – KI ist bereits im Einsatz

Viele Fachkräfte nutzen bereits ChatGPT (und ähnliches) informell.

32%

der Arbeitsstätten führen keine Gefährdungsbeurteilung durch.

Aktualität und Detailgrad dabei noch unberücksichtigt

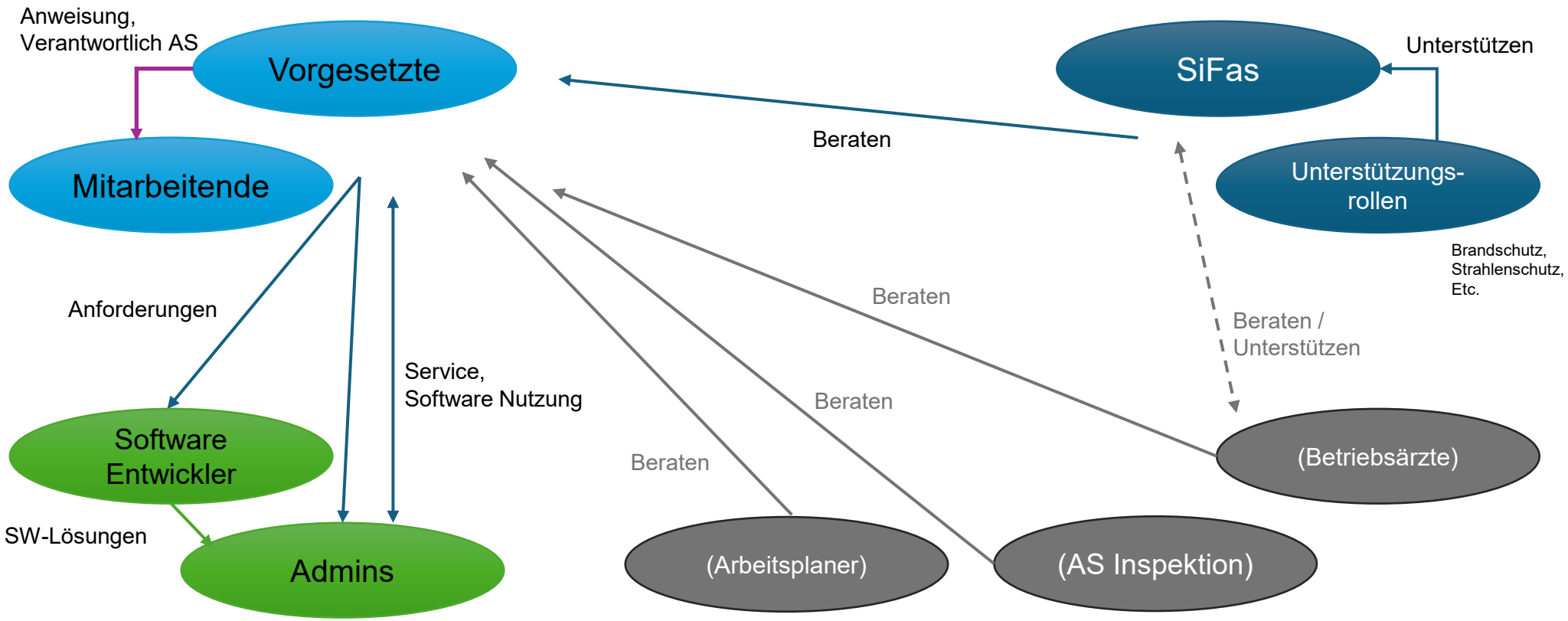
Problematik

- Informationssicherheit
- Halluzinationen
- Spezifikation
- Abbildung des Kontextes



Ungeregelte Schatten-IT

Arbeitsschutz: Ein komplexes soziotechnisches System



Von Anforderungen und Anwendungsszenarien



Initiale Anforderungsanalyse in Experteninterviews:

- Gegenwärtiger Prozess
- Abschätzung Veränderungen durch KI-Einsatz

➢ Anforderungen: Digitalisierungswünschen bis KI-spezifisches

Anwendungszenarien und Scenario-Based Design

- Expertenworkshops
- Gegenwärtige Prozesse (Problemszenario)
- Veränderung durch KI-Einsatz (Activity Szenario)

➢ Anwendungsszenarien, Positive / negative Eigenschaften



Einblick: Anforderungs-integrierte Szenarios

Mrs. Winkler looks at her watch and is surprised that its already near 12 o'clock. "Let's continue later on, the others are surely waiting for us to go to the cantina." Torsten closes all applications and shuts down his PC. "Was that smart now?", asks Mrs. Winkler. But Torsten waves her off: "We will get there again quickly."

And indeed, after their lunch break, Torsten requests help with their issue again and, although he didn't ask exactly the same as before, the SST arrives at the same conclusions.

F6

Problem: From similar basic workplace descriptions, I must arrive at similar conclusions even when I formulate it differently or add / subtract a insignificant details.

Requirement: As any role, the system's outputs must be robust enough to allow for differently styled descriptions of the same workplace to arrive at the same or at least comparable results.

Possible Solution: Calibrate the underlying LLM to a minimum viable temperature parameter.

The SST suggests that the provided data makes up for only about forty percent of the data relevant for this workplace. Accordingly, the SST estimates the confidence for its output at around a meager fifty percent. The SST proposes to add a list of the machines being used in the workshop. It furthermore offers to explain the identified risks in more detail as well as to highlight which parts of the input were relevant for generating its answer.

F11

Problem: If I rely on AI-hallucinated content, I could arrive at unfitting analysis and countermeasures, which in turn could lead to harm.

Requirement: As any user, the system always informs me about its confidence in its output, so I have a sense about how much I can rely on the generated content.

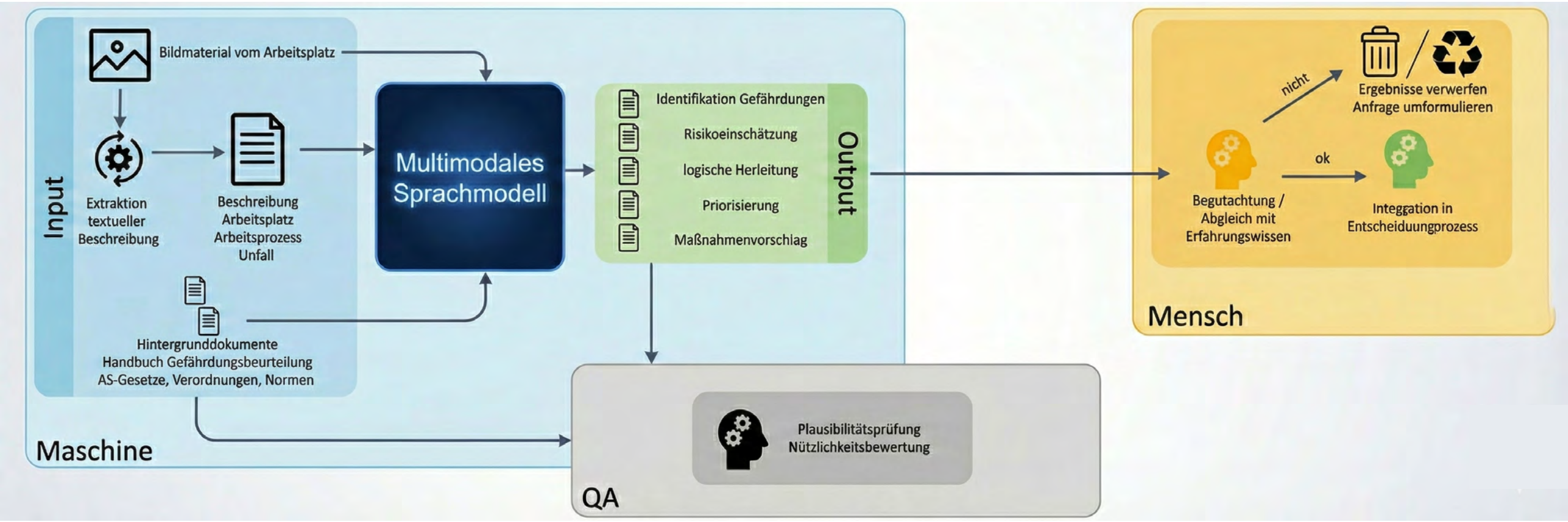
Possible Solution: See Uncertainty Quantification, e.g. Lin et al. [11].

Prototypsystem

zur Erprobung hinsichtlich Prompt-Design, Hintergrunddokumenten



Modelle kommen und gehen – Der Rahmen bleibt



Prototypsystem – Systeminstruktion – Leitende Prinzipien

zur Erprobung hinsichtlich Prompt-Design, Hintergrunddokumenten

- Systeminstruktion
  - U.a. Priorität Gesundheit und Sicherheit, Rechtskonformität, Praxisbezug, Aktualität, Neutralität und Objektivität
  - Auftreten: Selbstbezeichnung, Critiquing, Eigene und menschliche Aufgaben im Prozess
  - Spezifische Requirements (sofern von hier beeinflussbar)

- Hintergrunddokumente
  - Gesetze
  - Verordnungen
  - Technische Regeln
  - Handbücher & Empfehlungen
    - BAuA Handbuch Gefährdungsbeurteilung
    - LV59 Handlungsanleitung Zur Überprüfung der Gefährdungsbeurteilung
    - BMBFSFJ Leitfaden zum Mutterschutz
    - GDA Empfehlungen zur Berücksichtigung psychischer Belastung in der Gefährdungsbeurteilung

(Experten-) evaluation



- Wünschenswert: Umfassender Benchmark
  - Mit Experten ausgearbeitete Frage-Antwort-Paare
  - Score als objektives Gütemaß
  - Ermöglicht Vergleich: Modelle, Parametrisierungen, Menschl. Baseline
  - Aber: Hoher Aufwand
    - Teilgebiete gleichmäßig abdecken
    - Hunderte bis tausende Frage-Antwort-Paare
    - Dynamischer regulatorischer Kontext erfordert regelmäßige Updates
- Pragmatischer Ansatz: Expertenevaluation
  - Hier: Online (Videocall und Webinterfaces), n=28
  - Aufgabe: GBU-Skizzen erstellen
    - Vorgegebene Arbeitsplatzbeschreibung und Bebilderung
  - Messung
    - 30 Minuten Interaktion mit Think-Aloud-Protokoll
    - Entstandenen Skizzen (Reproduzierbarkeit der Ergebnisse)
    - Trust in Automation (Körber 2018)
    - User Experience Questionnaire (Laugwitz et al. 2008)
      - Fokus auf Pragmatische Qualität (Durchschaubarkeit, Effizienz, Vorhersagbarkeit)
    - 3 Iterationen über Prompt-Design und Hintergrunddokumente

Takeaway

1. KI ist nicht nur ChatGPT.
2. KI ist nicht immer die beste Lösung.
3. Um große Textmengen zu bewältigen vielleicht aber schon.
4. To Do: Evaluation hinbekommen.

Vielen Dank für Ihre Aufmerksamkeit!

Martin Westhoven

2.0 Wissenschaftliche Leitung,
Koordinierung KI-Forschung
Wissenschaftlicher Mitarbeiter

westhoven.martin@baua.bund.de

Arn Dietz

FG 2.4 – Künstliche Intelligenz in der Arbeitswelt
Wissenschaftlicher Mitarbeiter

dietz.arn@baua.bund.de

b a u a :
Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)
Fachbereich 2 – Produkte und Arbeitssysteme