Westhoven, Martin; Dietz, Arn

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

Zeitschrift für Arbeitswissenschaft

Provided in Cooperation with:
Springer Nature

Suggested Citation: Westhoven, Martin; Dietz, Arn (2025) : KI-Unterstützung zur Verarbeitung von Text- und Umweltdaten in der Gefährdungsbeurteilung, Zeitschrift für Arbeitswissenschaft, ISSN 2366-4681, Springer, Berlin, Heidelberg, Vol. 79, Iss. 3, pp. 451-459, https://doi.org/10.1007/s41449-025-00482-5

Zeitschrift für Arbeitswissenschaft (2025) 79:451–459
https://doi.org/10.1007/s41449-025-00482-5

PRAXISBEITRAG

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

Martin Westhoven[1] · **Arn Dietz**[1]

Angenommen: 4. September 2025 / Online publiziert: 24. September 2025

Zusammenfassung

Der Artikel untersucht, wie Künstliche Intelligenz (KI) die Gefährdungsbeurteilung im betrieblichen Arbeitsschutz verbessern kann. Trotz ihrer gesetzlichen Verpflichtung weisen viele Unternehmen, insbesondere kleine und mittlere, Defizite bei der Durchführung auf. Um diese Lücke zu schließen, werden zwei KI-gestützte Ansätze vorgestellt: Zum einen können mittels großer Sprachmodelle (LLMs) und der Retrieval-Augmented Generation (RAG)-Methode Systeme unstrukturierte Textdokumente (Gesetze, Unfallberichte) analysieren. Sie identifizieren relevante Gefahren und schlagen Maßnahmen vor, basierend auf vertrauenswürdigen, unternehmensspezifischen Daten. Dieser Ansatz reduziert „Halluzinationen" der KI und erhöht die Nachvollziehbarkeit. Zum anderen können günstige Sensoren kontinuierlich Umweltdaten wie Temperatur oder Luftqualität überwachen. Durch Anomalieerkennung mittels neuronaler Netze (Autoencoder) kann die KI subtile Abweichungen vom Normalzustand erkennen, die bei sporadischen Kontrollen oft übersehen werden. Die Autoren leiteten in Interviews und Workshops 17 spezifische Anforderungen an solche Systeme ab, die Transparenz, Nachvollziehbarkeit und die menschliche Entscheidungshoheit (Human-in-the-Loop-Prinzip) betonen. KI wird dabei als „Co-Pilot" verstanden, der Experten unterstützt, nicht ersetzt. Die finale Verantwortung verbleibt beim Menschen. Die sichere und vertrauenswürdige Integration von KI erfordert daher Erklärbarkeit und eine sorgfältige Gestaltung.

Schlüsselwörter Künstliche Intelligenz · Gefährdungsbeurteilung · Natural Language Processing · Sensordatenverarbeitung · Anomalieerkennung

AI-support for processing textual and environmental data in risk assessments

Abstract

The article examines how Artificial Intelligence (AI) can improve risk assessment in occupational safety. Despite being a legal obligation, many companies, especially small and medium-sized ones, show deficiencies in their implementation. To address this gap, two AI-supported approaches are presented: First, using large language models (LLMs) and the Retrieval-Augmented Generation (RAG) method, systems can analyze unstructured text documents (laws, accident reports). They identify relevant hazards and suggest measures based on trusted, company-specific data. This approach reduces AI "hallucinations" and increases traceability. Second, inexpensive sensors can continuously monitor environmental data such as temperature or air quality. Through anomaly detection using neural networks (autoencoders), the AI can recognize subtle deviations from the normal state that are often overlooked during sporadic checks. Based on interviews and workshops, the authors derived 17 specific requirements for such systems, emphasizing transparency, traceability, and human decision-making authority (the human-in-the-loop principle). AI is understood as a "co-pilot" that supports experts, not replaces them. The final responsibility remains with the human. Therefore, the safe and trustworthy integration of AI requires explainability and careful design.

✉ Martin Westhoven
westhoven.martin@baua.bund.de

Arn Dietz
dietz.arn@baua.bund.de

[1] Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, Friedrich-Henkel-Weg 1–25, 44149 Dortmund, Deutschland

Keywords Artificial intelligence · Risk assessment · Natural language processing · Sensor data processing · Anomaly detection

1 Einleitung

Künstliche Intelligenz (KI) entwickelt sich mit großer Geschwindigkeit zu einer transformativen Technologie, die zunehmend alle Sektoren der Arbeitswelt durchdringt. Während der Fokus der öffentlichen und wirtschaftlichen Debatte primär auf Anwendungen liegt, die direkte Optimierungspotenziale und kurzfristige ökonomische Vorteile versprechen, rücken die Potenziale für den Arbeits- und Gesundheitsschutz erst langsam in den Vordergrund. Da der Arbeitsschutz ökonomische Kennzahlen meist nur indirekt beeinflusst, erhalten entsprechende KI-Anwendungen vergleichsweise wenig Beachtung. Dabei bietet die technologische Dynamik gerade hier enorme Chancen, die Sicherheit und Gesundheit der Beschäftigten proaktiv zu gestalten und das hohe Niveau des Arbeitsschutzes in Deutschland weiter zu verbessern.

Ein zentrales Anwendungsfeld in den hier präsentierten Projekten ist die gesetzlich vorgeschriebene Gefährdungsbeurteilung, die als Kerninstrument des betrieblichen Arbeitsschutzes fungiert. Ihr primäres Ziel ist die systematische Identifikation, Bewertung und Minimierung von Gefährdungen für Beschäftigte am Arbeitsplatz. Die Gefährdungsbeurteilung selbst ist ein vielschichtiges und komplexes Feld, das sowohl fundiertes technisches Verständnis als auch soziale Kompetenzen im Umgang mit den Beschäftigten erfordert.

Die betriebliche Praxis zeigt jedoch weiterhin Defizite auf, trotz Verbesserungen gegenüber der Vergangenheit. Eine umfassende Betriebs- und Beschäftigtenbefragung aus den Jahren 2023 und 2024 (GDA-Arbeitsgruppe Evaluation 2025) zeigt entsprechend eine nach wie vor große Schutzlücke: Für rund 32 % der Arbeitsstätten in Deutschland werden keine Gefährdungsbeurteilungen durchgeführt. Insbesondere kleine und mittelständische Unternehmen (KMU) stehen hier vor großen Herausforderungen, die von mangelnden Ressourcen bis hin zu fehlendem Fachwissen reichen. Aber auch in Betrieben, in denen Gefährdungsbeurteilungen stattfinden, lässt sich aus der reinen Existenz keine Aussage über deren Qualität, Vollständigkeit oder Aktualität ableiten. Die Qualität hängt oft maßgeblich von der Erfahrung, der Kompetenz und dem Engagement der verantwortlichen Fachkraft ab.

Ziel dieses Beitrages ist es daher, zwei Anwendungen von KI zur Unterstützung der Gefährdungsbeurteilung systematisch zu erschließen und der betrieblichen Praxis sowohl die grundsätzliche Machbarkeit als auch konkrete Gestaltungsoptionen aufzuzeigen. Die vorgestellten Lösungsansätze fußen auf jüngsten Fortschritten im Bereich der KI (vgl. Vaswani et al. 2017; Chen et al. 2018), und leiten daraus praxisnahe Anforderungen für die Entwicklung und den operativen Einsatz solcher KI-Werkzeuge ab. Da Anwendungen zur Unterstützung von Gefährdungsbeurteilungen sicherheitsrelevant sind, ist eine außerordentlich sorgfältige Erhebung und Berücksichtigung von Anforderungen und Rahmenbedingungen erforderlich, um zu verhindern, dass die eingesetzten KI-Werkzeuge selbst neue Risiken generieren. Die hier vorgestellten Arbeiten sollen Praktikerinnen, Praktikern und Softwareentwicklungsunternehmen eine Orientierung bieten, wie KI-gestützte Werkzeuge für die Gefährdungsbeurteilung sicher und menschzentriert gestaltet werden können.

2 Die Gefährdungsbeurteilung in der Praxis: Komplexität und Unterstützungspotenziale

Der Arbeitsschutz in Deutschland verfolgt einen ganzheitlichen Ansatz, der nicht nur die reine Prävention von Unfällen, sondern auch den Schutz der Gesundheit und die menschengerechte Gestaltung der Arbeit umfasst. Die Gefährdungsbeurteilung ist das zentrale Instrument, um diesen Anspruch in die betriebliche Praxis zu überführen. Ihre Durchführung ist jedoch mit zahlreichen Herausforderungen verbunden, die die Potenziale für eine KI-gestützte Unterstützung verdeutlichen.

- *Subjektivität, Komplexität und Zielkonflikte:* Die individuelle Wahrnehmung der Arbeitsumwelt kann stark variieren. Als Beispiel sei die empfundene Raumtemperatur genannt, bei der eine Person 19 °C als ausreichend empfindet, während eine andere 21 °C als Minimum ansieht. Noch komplexer wird es beim Zusammenspiel verschiedener Umweltparameter, deren Implikationen in den Arbeitsstättenregeln nicht immer eindeutig geregelt sind, was zu Zielkonflikten führen kann. Ein anschauliches Beispiel ist die Verlegung von Dachpappe mittels Lötlampe. Dieser Prozess generiert erhebliche Hitze und erfordert das Tragen von Schutzkleidung. Bei hohen Außentemperaturen kann genau diese Schutzkleidung jedoch zu einer Überhitzung der beschäftigten Person führen, was den Einsatz leichterer Kleidung ratsamer erscheinen ließe. Die Analyse und Auflösung solcher Konflikte obliegt der Sicherheitsfachkraft und erfordert ein hohes Maß an Erfahrung.
- *Aufwand der Datenerfassung und -analyse:* Obwohl für die Überwachung von Umweltparametern wie Lärm, Kli-

ma oder Luftqualität standardisierte Messverfahren existieren, sind diese oft komplex und aufwendig in der Anwendung. Infolgedessen werden die relevanten Parameter häufig nur bei einem begründeten Verdacht auf eine Abweichung von den Vorgaben gemessen, anstatt präventiv und kontinuierlich.

- *Arbeitsplatzbegehung als Momentaufnahme:* Eine Arbeitsplatzbegehung, die üblicherweise zur Beurteilung herangezogen wird, ist immer nur eine Stichprobe des Arbeitsgeschehens. Der wesentliche Vorteil einer sensorbasierten Erfassung liegt darin, dass die kontinuierliche Präsenz der Sensoren auch die Erfassung seltener, flüchtiger Ereignisse ermöglicht. Solche Vorkommnisse bleiben bei der klassischen Methode oft unberücksichtigt, da die zuständigen Fachkräfte typischerweise ein hohes Arbeitsvolumen haben und zeitlich limitiert sind.

Folgende fiktive Beispiele illustrieren, wie eine kontinuierliche Datenerfassung durch Sensoren hier potenziell Fehler und Fehldeutungen vermeiden kann:

- Ein undichter Behälter mit flüchtigen Substanzen (z. B. Reinigungsalkohol) in einem Schrank wird von den Beschäftigten zwar geruchlich wahrgenommen, aber nicht als Sicherheitsrisiko eingeschätzt. Da der Schrank zum Zeitpunkt der Begehung geschlossen ist, bemerkt die Fachkraft für Arbeitssicherheit den Geruch nicht. Sensordaten hingegen würden zu den entsprechenden Zeiten einen Anstieg der flüchtigen organischen Substanzen in der Raumluft objektiv aufzeigen.
- Eine defekte Absaugung an einer Werkbank führt bei der Bearbeitung zu einer erhöhten Feinstaubkonzentration, die kontinuierlich von Sensoren registriert werden kann, während sie bei einer kurzen Begehung möglicherweise nicht auffällt.
- An einem heißen Sommertag steigt die Temperatur in einer Werkshalle auf über 45 °C. Die extern tätige Sicherheitsfachkraft ist an diesem Tag nicht vor Ort. Obwohl die zu hohen Temperaturen unter den Beschäftigten bekannt sind, wird dies aufgrund der betrieblichen Kultur nicht gemeldet. Die Sensordaten würden diesen kritischen Temperaturanstieg objektiv und unmissverständlich dokumentieren.

Genau an diesen Punkten kann KI ansetzen: Indem sie große Mengen unstrukturierter Textdaten analysiert, um Wissen zugänglich zu machen, und indem sie kontinuierlich Sensordaten verarbeitet, um Muster und Abweichungen zu erkennen, die dem menschlichen Beobachter verborgen bleiben.

3 Methodisches Vorgehen zur Anforderungsermittlung

Um praxistaugliche und akzeptierte KI-Anwendungen zu gestalten, ist ein tiefes und detailliertes Verständnis der realen Arbeitsprozesse sowie der Bedürfnisse der beteiligten Akteure unerlässlich. Aus diesem Grund wurde ein zweistufiges, qualitatives Vorgehen zur systematischen Anforderungsermittlung gewählt, das auf etablierten Methoden der Software-Ergonomie und der qualitativen Forschung basiert.

3.1 Experteninterviews

Zwischen 2022 und 2023 wurden insgesamt 15 semi-strukturierte Experteninterviews durchgeführt. Interviews gelten als eine der effektivsten Methoden zur Anforderungserhebung (Davis et al. 2006), da sie eine Rekonstruktion des Expertenwissens zu einem spezifischen Thema ermöglichen (Pfadenhauer 2009). Es ist jedoch zu beachten, dass Experten ihr Handeln in routinierten Abläufen selten explizit reflektieren, was zu Auslassungen führen kann. Zudem liegt Expertenwissen oft implizit vor, was seine Verbalisierung erschwert (Preim und Dachselt 2010). Als Zielgruppe wurden gezielt Fachkräfte für Arbeitssicherheit gewählt, da diese über das detaillierte Prozesswissen verfügen und als zentrale Ansprechpartner die Anforderungen weiterer involvierter Akteure kennen. Die ersten sechs Interviews fanden mit internen Fachkräften der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin statt, gefolgt von neun weiteren Interviews mit Vertreterinnen und Vertretern unterschiedlicher Unternehmen des produzierenden Gewerbes sowie eines externen Arbeitsschutz-Dienstleisters. Tiefergehende Ergebnisse zu den Interviewergebnissen sind in Westhoven (2022) publiziert.

3.2 Szenarienworkshops

Ergänzend zu den Interviews wurden zwei Szenarienworkshops durchgeführt. Diese basieren auf dem Konzept des Scenario-Based Design (SBD), das die Erstellung konkreter Nutzungsszenarien für ein System in den Mittelpunkt stellt (Rosson und Carroll 2002). Ein zentraler Vorteil von Szenarien ist ihre Funktion als leicht verständliches Kommunikationsmedium zwischen verschiedenen Stakeholdern. In den Workshops wurde zunächst der Status quo eines Aufgabenprozesses als „Problemszenario“ abgebildet. In der anschließenden Designphase wurden diese in „Aktivitätsszenarien“ überführt, die den Einsatz der zu entwickelnden Technologie und die resultierenden Prozessveränderungen beleuchten. Die finalen, in den Workshops erarbeiteten Szenarien wurden mithilfe eines Chatbots (Modell: Google Gemini 1.5 Flash) weiter ausgearbeitet, um die rohen Ideen in

detaillierte, narrative Geschichten zu überführen, die anschließend auf Plausibilität geprüft wurden. Eine detaillierte Beschreibung der Szenarienworkshops und der Anforderungsintegration findet sich in Westhoven und Herrmann (2025).

Aus den umfangreichen qualitativen Daten der Interviews und Workshops wurden zunächst allgemeine Anforderungen abgeleitet und anschließend nach ihrer Spezifität für KI-Anwendungen gefiltert. Insgesamt konnten so 17 KI-spezifische Anforderungen identifiziert werden, die im Folgenden detailliert vorgestellt und erläutert werden.

4 KI-gestützte Lösungsansätze für die betriebliche Praxis

Aufbauend auf der Anforderungsanalyse wurden zwei technologische Ansätze vertieft, die auf die zentralen Herausforderungen der Gefährdungsbeurteilung abzielen: die Verarbeitung von Textdaten zur Wissensextraktion und die Auswertung von Sensordaten zur Umgebungsüberwachung.

4.1 Textdaten intelligent erschließen: Unterstützung durch Große Sprachmodelle (LLMs)

Ein Großteil des relevanten Wissens im Arbeitsschutz liegt in unstrukturierter, textbasierter Form vor: in Gesetzen, technischen Regeln, Normen, früheren Gefährdungsbeurteilungen und Unfallberichten. Die Veröffentlichung von leistungsfähigen Großen Sprachmodellen (LLMs) wie GPT-4 (Achiam et al. 2023) hat die Vision einer KI realisiert, die auf Basis einer Arbeitsplatzbeschreibung Gefährdungen identifizieren und Gegenmaßnahmen vorschlagen kann. Anstatt ein eigenes, konkurrierendes Modell von Grund auf zu entwickeln, was enorme Mengen an zugänglichen Trainingsdaten erfordern würde und im aktuellen Entwicklungstempo zudem schnell wieder veraltet wäre, konzentriert sich der vorgestellte Ansatz auf die Methode der Retrieval-Augmented Generation (RAG), die von Lewis et al. (2020) eingeführt wurde.

Für Praktikerinnen und Praktiker lässt sich der RAG-Ansatz wie folgt erklären: Stellt eine Fachkraft eine Anfrage an das System (z. B. „Welche Gefährdungen bestehen bei Schweißarbeiten in engen Räumen und welche Schutzmaßnahmensind erforderlich?“), durchsucht das System nicht nur sein allgemeines, vortrainiertes Wissen. Stattdessen führt es parallel eine gezielte semantische Suche in einer hinterlegten, vertrauenswürdigen und domänenspezifischen Wissensdatenbank durch. Diese Datenbank kann unternehmensspezifische Richtlinien, alle bisherigen Gefährdungsbeurteilungen des Betriebs oder relevante Auszüge aus dem Vorschriften- und Regelwerk enthalten. Die relevantesten gefundenen Dokumente oder Textpassagen werden der ursprünglichen Anfrage als zusätzlicher, fundierter Kontext beigefügt, bevor das Sprachmodell eine Antwort formuliert und generiert. Der schematische Aufbau eines solchen Systems ist in Abb. 1 illustriert. Der Ansatz bietet entscheidende Vorteile für die Praxis:

- Verbesserte Qualität und Spezifität: Die Antworten sind spezifischer und basieren auf geprüften, für den Anwendungsfall relevanten Quellen, anstatt auf allgemeinem Weltwissen, das veraltet oder ungenau sein kann.
- Reduzierung von Halluzinationen: Das Risiko von KI-Halluzinationen, also frei erfundenen, aber plausibel

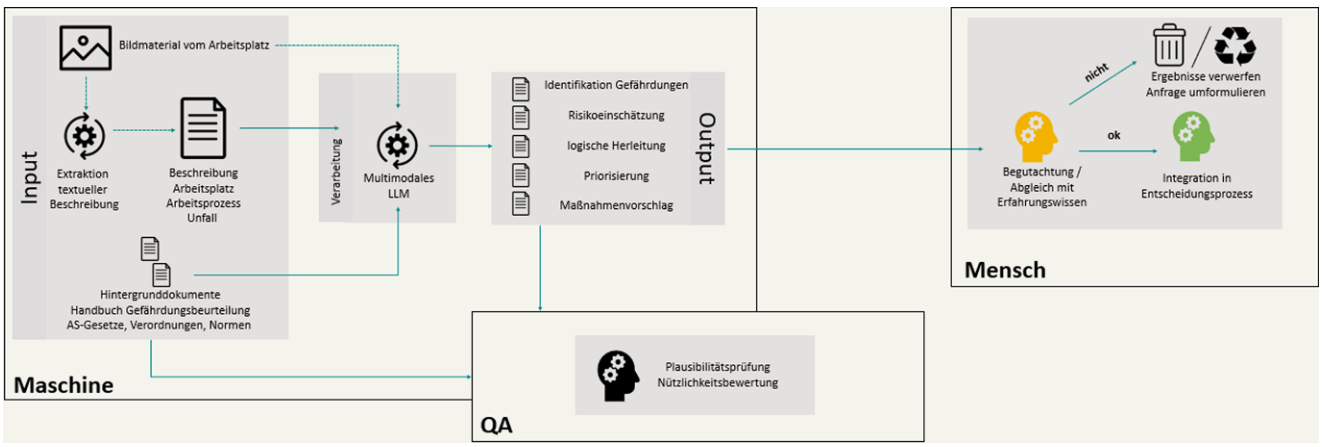

Abb. 1 Schematischer Aufbau eines RAG-Systems zur Unterstützung von Gefährdungsbeurteilungen. Die Nutzeranfrage wird durch relevante Informationen aus einer internen Wissensdatenbank (z. B. Gesetze, frühere GBUen) angereichert, bevor das Sprachmodell die kontextualisierte Antwort generiert

Fig. 1 Schematic diagram of a RAG (Retrieval-Augmented Generation) system to support risk assessments. The user query is enriched with relevant information from an internal knowledge base (e.g., laws, previous risk assessments) before the language model generates the contextualized response

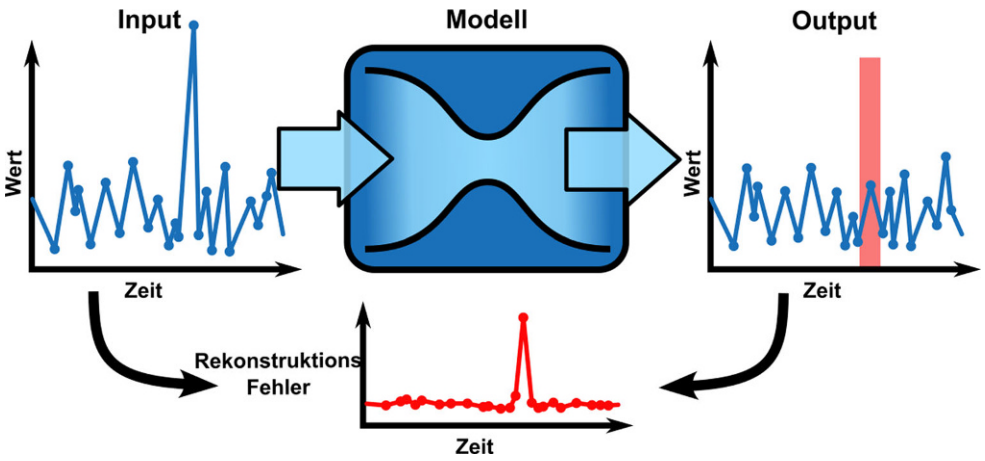

Abb. 2 Schematische Darstellung eines Autoencoders zur Anomaliedetektion. Ein Input-Datenstrom, der eine Anomalie enthält (hier ein Ausreißerwert), wird durch das auf „normale" Daten trainierte Modell geleitet. Der Output wird ohne die Anomalie rekonstruiert, da das Netz gelernt hat, den Normalzustand abzubilden. Die Differenz zwischen Input und Output (der Rekonstruktionsfehler) macht die Anomalie quantifizierbar und sichtbar. (Abbildung nach Baudzus 2024)

Fig. 2 Schematic diagram of an autoencoder for anomaly detection. An input data stream containing an anomaly (in this case, an outlier) is processed by the model, which has been trained on "normal" data. The output is reconstructed without the anomaly, as the network has learned to represent the normal state. The difference between the input and output (the reconstruction error) renders the anomaly quantifiable and visible. (Figure adapted from Baudzus 2024)

klingenden Informationen (siehe Zhang et al. 2023), wird signifikant reduziert, da die Antwort im bereitgestellten Kontext verankert wird.

- Nachvollziehbarkeit und Transparenz: Das System kann seine Quellen direkt zitieren, was die Nachvollziehbarkeit und Überprüfbarkeit der Ergebnisse massiv erhöht – eine zentrale Anforderung im sicherheitskritischen Kontext des Arbeitsschutzes.

4.2 Umweltdaten kontinuierlich überwachen: Anomalieerkennung mittels Sensorik und KI

Der zweite Lösungsansatz widmet sich der Überwachung der physischen Arbeitsumgebung. Dank des technischen Fortschritts in der Elektronikproduktion sind heute kostengünstige und einfach zu installierende Sensorprodukte verfügbar, die Umweltparameter wie Temperatur, Luftfeuchtigkeit, Lärmpegel oder die Konzentration flüchtiger organischer Verbindungen (VOCs) mit einer für qualitative Analysen ausreichenden Qualität erfassen können. Auch wenn diese Sensoren nicht die von den Arbeitsstättenregeln vorgegebenen Gütemerkmale erfüllen, können ihre Daten wertvolle Informationen liefern. Die eigentliche Herausforderung liegt darin, aus den kontinuierlich anfallenden, riesigen Zeitreihendatenmengen relevante und handlungsleitende Informationen zu gewinnen.

Hier kommt die KI-gestützte Zeitreihenanomaliedetektion ins Spiel (siehe Shaukat et al. 2021 für einen Überblick). Anstatt nur starre, vordefinierte Grenzwerte zu überwachen, wie sie etwa in den Arbeitsstättenregeln definiert sind, kann ein KI-Modell lernen, wie der komplexe, dynamische „Normalzustand" an einem spezifischen Arbeitsplatz aussieht. Eine vielversprechende Methode hierfür ist die rekonstruktionsbasierte Anomaliedetektion unter Verwendung eines neuronalen Netzes mit einer sogenannten Autoencoder-Struktur (beschrieben von Malhotra et al. 2016, siehe Abb. 2). Im Folgenden wird kurz auf die Funktion eines Autoencoders als Anomaliedetektor eingegangen: Das Modell wird mit einer großen Menge an Sensordaten aus dem normalen Betriebsablauf trainiert. Es lernt dabei, die typischen Muster und Zusammenhänge der verschiedenen Datenströme zu komprimieren und anschließend wieder möglichst exakt zu rekonstruieren. Um sicherzustellen, dass das Modell nicht generell lernt, alle möglichen Inputs zu rekonstruieren, beinhaltet es einen Informationsflaschenhals, der diese Möglichkeit ausschließt. Das Modell muss also bei jeder Rekonstruktion die Eigenschaften des Inputs auswählen, die für die Rekonstruktion relevant sind. Das Netzwerk wird somit zur Referenz für den Normalzustand. Wenn nun ein ungewöhnliches, seltenes Ereignis auftritt, eine plötzliche Hitzeentwicklung, ein unerwarteter Anstieg von Substanzen in der Luft oder eine ungewöhnliche Lärmspitze, scheitert das System an der perfekten Rekonstruktion dieser „unnormalen" Daten, da es diese Muster im Training nie gelernt hat. Dieser „Rekonstruktionsfehler", also die Differenz zwischen dem originalen Input und dem rekonstruierten Output, wird automatisch erkannt und als Anomalie gemeldet.

Der praktische Nutzen ist enorm: Das System kann nicht nur bekannte Grenzwertüberschreitungen melden, sondern auch subtile, unbekannte Abweichungen vom erlernten

Normalzustand aufdecken, die auf eine bisher unerkannte, schleichende oder seltene Gefährdung hindeuten könnten.

5 KI-bezogene Anforderungen an praxistaugliche Unterstützungssysteme

Aus den Interviews und Workshops wurden 17 spezifische Anforderungen abgeleitet, die ein KI-System erfüllen muss, um im hochgradig soziotechnischen und sicherheitskritischen Kontext des Arbeitsschutzes sicher und nützlich zu sein. Diese Anforderungen gehen über allgemeine Aspekte der Digitalisierung hinaus (siehe hierzu Westhoven 2022) und werden zur besseren Übersicht in drei zentrale Bereiche gegliedert:

5.1 Anforderungen an Dateneingabe, Wissensbasis und Analysefähigkeit

- *Multimodale Eingabe (F1):* Eine rein textuelle Beschreibung von Arbeitsplätzen kann mühsam oder unmöglich sein. Das System sollte daher für jede Benutzerrolle verschiedene, leicht verfügbare Medien wie Bilder, Videos oder Audioaufnahmen als Eingabe akzeptieren und relevante Informationen daraus extrahieren können.
- *Geführte und kontextsensitive Datenerfassung (F2):* Die Bereitstellung unzureichender oder irrelevanter Daten kann zu falschen Analysen führen. Das System sollte daher, basierend auf einer ersten Eingabe, den Arbeitsplatztyp erkennen (z. B. über semantische Suche) und aktiv nach weiteren, für diesen Arbeitsplatz typischen und relevanten Parametern fragen, die in der bisherigen Beschreibung noch fehlen.
- *Integration von externem und internem Hintergrundwissen (F3):* Jede Gefährdungsbeurteilung basiert auf einem großen Korpus an allgemeinem Arbeitsschutzwissen. Das System muss in der Lage sein, auf dieses übergeordnete Wissen (Gesetze, Regeln, Standards, Unternehmensrichtlinien) zuzugreifen und es automatisch mit der spezifischen Arbeitsplatzbeschreibung zu verknüpfen (z. B. durch semantische Suche).
- *Vergleich mit bestehenden Fällen zur Konsistenzsicherung (F4):* Der manuelle Vergleich mit allen bestehenden Gefährdungsbeurteilungen in einem Unternehmen ist oft undurchführbar. Das System soll für Vorgesetzte oder Sicherheitsfachkräfte automatisch passende bestehende Beurteilungen (oder Teile davon) finden und ausgeben können, um die Konsistenz zu erhöhen und redundante Arbeit zu vermeiden.
- *Identifikation relevanter Umgebungsfaktoren (F5):* Oft ist unklar, welche Umgebungsfaktoren für einen Arbeitsplatz überhaupt relevant sind. Das System sollte eine Übersicht und Begründung zu wichtigen Parametern geben, die den zu begutachtenden Arbeitsplatz beeinflussen, indem es aus bestehenden Beurteilungen und Hintergrunddokumenten abschätzt, welche Faktoren vorherrschend sind.

5.2 Anforderungen an Transparenz, Nachvollziehbarkeit und Verlässlichkeit der Ergebnisse

- *Robuste und reproduzierbare Ergebnisse (F6):* Aus ähnlich formulierten Beschreibungen desselben Arbeitsplatzes müssen zu ähnlichen Schlussfolgerungen führen, auch wenn unwesentliche Details variieren. Die Ausgaben des Systems müssen robust genug sein, um dies zu gewährleisten.
- *Interpretation vage formulierter Regeln (F7):* Gesetze und technische Regeln sind oft vage formuliert, was zu Unsicherheit bei der Umsetzung führt. Das System sollte Sicherheitsfachkräften detaillierte Erklärungen oder beispielhafte Gefährdungsbeurteilungen bereitstellen, die das Problem behandeln, um ein Gefühl dafür zu vermitteln, was und wie viel getan werden muss.
- *Aufzeigen von Wechselwirkungen und Nebenfolgen (F8):* Insbesondere Wechselwirkungen zwischen verschiedenen Merkmalen eines Arbeitsplatzes können leicht übersehen werden. Die Wissensbasis des Systems sollte in der Lage sein, das Verständnis der Nutzenden mit Informationen zu ergänzen, die aus den vorhandenen Daten ableitbar sind, z. B. durch logische Inferenz.
- *Begründung von vergangenen Entscheidungen (F9, F10):* Um die Schlüssigkeit von Argumentationslinien zu überprüfen und aus Gründen der Verantwortlichkeit, müssen Nutzende verstehen, warum frühere Entscheidungen über Gegenmaßnahmen getroffen wurden. Das System muss die Begründung für umgesetzte (oder auch verworfene) Maßnahmen dokumentieren und auf Anfrage präsentieren können.
- *Quantifizierung von Unsicherheit und Umgang mit Halluzinationen (F11):* Wenn sich Nutzende auf KI-halluzinierte Inhalte verlassen, kann dies zu unpassenden Gegenmaßnahmen und Schaden führen. Das System muss daher immer über seine Konfidenz in die eigene Ausgabe informieren, damit die Nutzenden einschätzen können, wie sehr sie dem generierten Inhalt vertrauen können (siehe auch Lin et al. 2023).
- *Schutz vor übermäßigem Vertrauen (Automation Bias) (F12):* Nutzende könnten dazu neigen, Systemvorschläge einfach zu akzeptieren, ohne sie richtig zu überprüfen. Das System sollte Anzeichen für ein solches Verhalten erkennen (z. B. durch Analyse der Reaktionszeit) und die Nutzenden informieren, damit sie die Ergebnisse bei Bedarf gründlicher überprüfen.

- *Hinweis auf Abweichungen von Standards (F13):* Abweichungen von Standard-Gegenmaßnahmen können zu suboptimalen Ergebnissen führen. Das System sollte Vorgesetzte oder Sicherheitsfachkräfte dabei unterstützen, Abweichungen von früheren Entscheidungen oder Standardverfahren zu bemerken, indem es verwandte Fälle sucht und auf konkrete Unterschiede hinweist.
- *Erklärbarkeit der Ergebnisrelevanz (F14):* Nutzende müssen verstehen, welche Teile ihrer Eingabe für die Ergebnisse des Systems relevant waren. Das System sollte relevante Textteile, Bildbereiche oder Audiopassagen hervorheben, um dies transparent zu machen.

5.3 Anforderungen an die Nutzerkontrolle, Verantwortung und Interaktion

- *Menschliche Entscheidungshoheit (Human-in-the-Loop) (F15):* Eine vollautomatische KI-Gefährdungsbeurteilung, bei der die Verantwortung diffus wird, ist inakzeptabel. Alle endgültigen Entscheidungen müssen von Menschen getroffen und beispielsweise durch (biometrische) Signaturen bestätigt werden.
- *Möglichkeit zum Überschreiben und Anpassen (F16):* Wenn Nutzende unerkannte Risiken oder bessere Maßnahmen sehen, müssen sie in der Lage sein, die Analyse und die Vorschläge des Systems zu überschreiben oder anzupassen. Das System muss hierfür einfache Möglichkeiten, z. B. via Prompting oder manuelle Bearbeitung, bereitstellen.
- *Interaktive Qualitätssicherung und Abschluss (F17):* Eine mangelhafte, automatisch generierte Dokumentation kann im Schadensfall zu erheblichen Problemen führen. Das System muss daher einen interaktiven Dialog zur Qualitätssicherung anbieten, um sicherzustellen, dass ein Mensch die Dokumentation vor dem endgültigen Abschluss überprüft hat.

6 Diskussion, Herausforderungen und Implikationen für die Praxis

Die Implementierung eines KI-gestützten Unterstützungssystems im hochgradig komplexen soziotechnischen System des betrieblichen Arbeitsschutzes ist mit erheblichen Herausforderungen verbunden. Die hier vorgestellten, allgemeinen Anforderungen können als wichtige Leitplanken für die Entwicklung realer Systeme dienen. Sie müssen jedoch stets durch branchen- und firmenspezifische Anforderungen ergänzt werden, da eine vollständige Generalisierbarkeit einer Einzellösung nicht gewährleistet werden kann. Für Praktikerinnen und Praktiker ist es entscheidend, die Grenzen der aktuellen Technologie realistisch einzuschätzen:

- *Die Herausforderung des „Drifts“ bei der Anomalieerkennung:* Ein zentrales Problem bei der sensordatenbasierten Analyse ist das Phänomen des Drifts. Der Trainingsprozess eines neuronalen Netzes ist rechenintensiv und wird meist nur einmalig durchgeführt. Das trainierte Netz bildet danach den Referenzzustand für „normal“ ab. Ein Drift tritt ein, wenn sich dieser Normalzustand dauerhaft ändert – beispielsweise durch die Installation einer neuen Klimaanlage, die das Umgebungsklima in einer Werkshalle verändert. Das auf den alten Zustand trainierte KI-System würde dann fälschlicherweise dauerhaft Anomalien melden. Ideale KI-Systeme benötigen daher einen Mechanismus, um solche Veränderungen zu erkennen und das Modell bei Bedarf erneut zu trainieren. Die Entwicklung solcher Mechanismen ist Gegenstand aktueller Forschung.
- *Keine Garantie für Korrektheit und die Notwendigkeit kritischer Prüfung:* Die Interpretation der Ergebnisse eines KI-Modells ist komplex, und es ist derzeit nicht möglich, Garantien für die Qualität der Ergebnisse zu geben. Insbesondere LLMs neigen trotz überzeugender Formulierungen zu Halluzinationen (siehe Huang et al. 2025 für einen Überblick) und einem mitunter übermäßig selbstsicheren Tonfall (siehe z. B. Mirza et al. 2025). RAG-Systeme reduzieren dieses Risiko, eliminieren es aber nicht. Die Ergebnisse müssen daher von der Fachkraft stets kritisch hinterfragt und verifiziert werden.
- *Expertenwerkzeug, kein Expertenersatz:* Die hier beschriebenen Systeme sind explizit als Werkzeuge für Expertinnen und Experten konzipiert. Sie sind nicht dazu gedacht, Aspekte der Gefährdungsbeurteilung zu übernehmen oder zu ersetzen, sondern Fachkräfte mit nützlichen, aufbereiteten Informationen bei ihren Aufgaben zu unterstützen. Da sich die Technologie nach wie vor in einem Entwicklungsstadium befindet, bedarf die Frage, ob und wie solche Systeme künftig auch Laien bei sicherheitskritischen Fragestellungen assistieren können, weiterer, intensiver Forschung.

Für die betriebliche Praxis bedeutet dies: KI sollte als ein intelligenter Co-Pilot verstanden werden, der auf verborgene Gefahren hinweist, relevantes Wissen bündelt und die aufwendige Dokumentation erleichtert. Die Verantwortung für die finale Beurteilung, die Abwägung von Maßnahmen und die endgültige Entscheidung verbleibt jedoch uneingeschränkt beim Menschen (Human-in-the-Loop-Prinzip). Die Nachvollziehbarkeit, Erklärbarkeit und Nachverfolgbarkeit der KI-generierten Ergebnisse sind die entscheidenden Kriterien für eine sinnvolle, vertrauenswürdige und sichere Integration von KI in die komplexen Mensch-Maschine-Systeme des Arbeitsschutzes.

7 Zusammenfassung und Ausblick

Dieser Artikel hat den erheblichen Unterstützungsbedarf im Rahmen des Gefährdungsbeurteilungsprozesses dargelegt und den Einsatz von Künstlicher Intelligenz in diesem Kontext begründet. Basierend auf einer fundierten qualitativen Erhebung mittels Experteninterviews und Szenarienworkshops wurden Anforderungen und Testszenarien entwickelt. Diese beleuchten zwei zentrale Forschungsfelder: Erstens die Nutzung von Large Language Models (LLMs) zur Auswertung von Arbeitsplatz- und Unfallbeschreibungen, und zweitens die Möglichkeiten der Sensordatenauswertung von Arbeitsumgebungsparametern am Beispiel der Anomalieerkennung.

Während die Anforderungserhebung und Szenarienentwicklung weitgehend abgeschlossen sind, befinden sich die Aspekte des Prompt Designs, also der optimalen Formulierung von Anfragen an LLMs, und des Benchmarkings zur objektiven Leistungsbewertung von LLMs für die Domäne des Arbeitsschutzes noch in der Bearbeitung. Zur Bewältigung des Drifts im Anomalieerkennungssystem ist der Einsatz einer Ensemble-Methode geplant, dem sogenannten „Bag of Models"-Ansatz, bei dem mehrere neuronale Netze parallel arbeiten und sukzessive durch neu trainierte Modelle ersetzt werden.

Zusammenfassend lässt sich festhalten, dass über die reine Funktionalität eines Unterstützungssystems hinaus zahlreiche weitere Rahmenbedingungen zu berücksichtigen sind, da der Arbeitsschutz ein komplexes soziotechnisches System darstellt. Insbesondere die Nachvollziehbarkeit, Erklärbarkeit und Nachverfolgbarkeit der KI-generierten Ergebnisse müssen äußerst sorgfältig konzipiert werden, um eine sinnvolle und sichere Integration von KI in ein solch komplexes Mensch-Maschine-System zu gewährleisten und das Vertrauen der Nutzenden zu gewinnen.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt M. Westhoven und A. Dietz geben an, dass kein Interessenkonflikt besteht.

Literatur

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, McGrew B (2023) Gpt-4 technical report. arXiv preprint arXiv:2303.08774

Amatriain X (2024) Prompt design and engineering: introduction and advanced methods. arXiv preprint arXiv:2401.14423

Baudzus A Entwicklung eines Assistenzsystems zur Gefährdungsbeurteilung unter Verwendung von Internet-of-Things-Technologien und Algorithmik aus dem Bereich des maschinellen Lernens und der Data Science. 70. Frühjahreskongress der Gesellschaft für Arbeitswissenschaft.

Chen J, Lin H, Han X, Sun L (2024) Benchmarking large language models in retrieval-augmented generation. Proc Aaai Conf Artif Intell 38(16):17754–17762

Chen Z, Yeo CK, Lee BS, Lau CT (2018) Autoencoder-based network anomaly detection. In: 2018 Wireless telecommunications symposium (WTS). IEEE, S 1–5

Chowdhary KR (2012) Natural language processing. MBM Engineering College, Jodhpur, India, lecture notes

Davis A, Dieste O, Hickey A, Juristo N, Moreno AM (2006) Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review. In: 14th IEEE International Requirements Engineering Conference (RE'06). IEEE, S 179–188

GDA-Arbeitsgruppe Evaluation (2025) Bericht zur Betriebs- und Beschäftigtenbefragung 2023/24

Hauer F, Pretschner A, Holzmüller B (2020) Re-using concrete test scenarios generally is a bad idea. In: 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, S 1305–1310

Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Liu T (2025) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 43(2):1–55

Kaner CJD (2013) An introduction to scenario testing. Florida Institute of Technology, Melbourne, S 1–13

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst 33:9459–9474

Li B, Jentsch C, Müller E (2023) Prototypes as explanation for time series anomaly detection. arXiv preprint arXiv:2307.01601

Lin Z, Trivedi S, Sun J (2023) Generating with confidence: Uncertainty quantification for black-box large language models. arXiv preprint arXiv:2305.19187

Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P, Shroff G (2016) LSTM-based encoder-decoder for multi-sensor anomaly detection. arXiv preprint arXiv:1607.00148

Mirza A, Alampara N, Kunchapu S, Ríos-García M, Emoekabu B, Krishnan A, Jablonka KM (2025) A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. Nat Chem: 1–8

Najafi B, Moaveninejad S, Rinaldi F (2018) Data analytics for energy disaggregation: Methods and applications. In: Big data application in power systems. Elsevier, S 377–408

Pfadenhauer M (2009) Das Experteninterview: Ein Gespräch auf gleicher Augenhöhe. In: Qualitative Marktforschung: Konzepte – Methoden – Analysen, S 449–461

Preim B, Dachselt R (2010) grundlagen, graphical user interfaces, informationsvisualisierung. Interaktive systeme, Bd. 1. Springer

Rosson MB, Carroll JM (2002) Usability engineering: scenario-based development of human-computer interaction. Morgan Kaufmann

Sclar M, Choi Y, Tsvetkov Y, Suhr A (2023) Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting. arXiv preprint arXiv:2310.11324

Shaukat K, Alam TM, Luo S, Shabbir S, Hameed IA, Li J, Javed U (2021) A review of time-series anomaly detection techniques: a step to future perspectives. In: Advances in information and communication: proceedings of the 2021 future of information and communication conference (FICC), Bd. 1. Springer, S 865–877

Strauss A, Corbin J (1998) Basics of qualitative research techniques

Sun J, Liao QV, Muller M, Agarwal M, Houde S, Talamadupula K, Weisz JD (2022) Investigating explainability of generative AI for code through scenario-based design. In: Proceedings of the 27th International Conference on Intelligent User Interfaces, S 212–228

Tonmoy SMTI, Zaman SM, Jain V, Rani A, Rawte V, Chadha A, Das A (2024) A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313,6

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:

Westhoven M (2022) Requirements for AI support in occupational safety risk analysis. In: Proceedings of Mensch und Computer 2022, S 561–565

Westhoven M, Herrmann T (2025) Scenarios and requirements for supporting workplace risk assessment with artificial intelligence. In: Artificial Intelligence in HCI

Zamanzadeh Darban Z, Webb GI, Pan S, Aggarwal C, Salehi M (2024) Deep learning for time series anomaly detection: a survey. ACM Comput Surv 57(1):1–42

Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Shi S (2023) Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219