Potenziale Künstlicher Intelligenz zur Risikoanalyse im betrieblichen Arbeitsschutz

Martin WESTHOVEN

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin,
Friedrich-Henkel-Weg 1-25, D-44149 Dortmund.

Kurzfassung: Die Dissertation beleuchtet Potenziale von KI-Methoden zur Unterstützung von Risikoanalysen im betrieblichen Arbeitsschutz. Dies bringt auf Datenebene verschiedene Probleme mit sich, die mit passenden Methoden und Algorithmen angegangen werden müssen. Zu nennen sind hierbei heterogene und unvollständige Daten sowie ein Mangel an gelabelten Daten für überwachtes Lernen. Darüber hinaus ist insbesondere auch die Interaktionsgestaltung zwischen Mensch und KI von hoher Bedeutung, da das Ergebnis einer Risikoanalyse im betrieblichen Arbeitsschutz unmittelbar sicherheitsrelevant ist und Fehler soweit wie möglich ausgeschlossen werden müssen.

Schlüsselwörter: Risikoanalyse, Gefährdungsbeurteilung, Künstliche Intelligenz, Interaktionsdesign

1. Problem- und Zielstellung

In diesem Vorhaben sollen die Möglichkeiten der Anwendung Künstlicher Intelligenz als Werkzeug des betrieblichen Arbeitsschutzes untersucht werden. Als Anwendungsfeld soll das Risikomanagement im betrieblichen Arbeitsschutz in den Fokus genommen werden. Hier entstehen heterogene und komplexe Daten, die sich nicht einfach hinsichtlich ihrer Zusammenhänge analysieren lassen. Künstliche Intelligenz liefert Ansätze, um diese Datenproblematik anzugehen und Zusammenhänge zu analysieren, zu modellieren und Risiken zu identifizieren.

Aufgrund der Vielfalt unterschiedlicher Risiken im Betrieb wird auf ein prototypisches System fokussiert, welches mit im Betrieb vorliegenden oder einfach zu erhebenden Daten abbildbar ist. Im Arbeitsschutz ergeben sich solche Daten z. B. aus den typischen Faktoren von Gefährdungsbeurteilungen, können aber auch abhängig vom Kontext Branchen-, bzw. Betriebseigenschaften sowie die immer häufiger vernetzt vorliegenden Maschinendaten umfassen. Hinsichtlich betrachteter Gefährdungsfaktoren werden physikalische Gefährdungen als Hauptaugenmerk bei der Vermeidung schwerer Unfälle in den Fokus genommen.

Die sehr erfolgreiche KI-Methode des Deep Learning kann das Vorhandensein einer Fülle von gelabelten Daten benötigen (siehe z. B. Shams (2014)), während im Risikomanagement aufgrund der Komplexität der Problemstellung nur spärliche objektive Daten sowie entsprechende Label zu erwarten sind (Clemen & Winkler 1999). Durch Einbindung von a priori-Wissen können jedoch sogenannte Few-Shot Learning-Ansätze zum Einsatz gebracht werden (siehe z. B. Wang et al. 2020), welche darauf zielen mit nur wenigen oder sogar nur einem Trainingsbeispiel zu arbeiten.

Um das Problem fehlender Daten anzugehen sei weiter das sogenannte Deep

Latent Variable Modelling (Kingma & Welling 2013) erwähnt, welches mittels eines generativen Ansatzes fehlende Daten imputiert und zuletzt in vielen Bereichen erfolgreich eingesetzt wurde.

Aufgrund der Komplexität der Aufgabe und der menschlichen Entscheidungshoheit soll der Ansatz hybrider Intelligenz (Dellermann et al. 2019) Beachtung finden. Eine Voraussetzung dafür ist Erklärbarkeit (Explainable AI, siehe Doran et al. 2017). Hybride Intelligenz erfordert weiter auch ein sorgfältig gestaltetes User Interface, ganz besonders, wenn der Systemoutput sicherheitskritisch ist. Dazu stellen bspw. Schmidt und Herrmann (2017) eine Anpassung von Shneiderman’s Golden Rules (Shneiderman et al. 2016) zur Diskussion.

Um Anforderungen zu ermitteln, soll für Experteninterviews mit einem größeren Unternehmen kooperiert werden. Der Gesprächsleitfaden greift auf das Konzept der Grounded Theory (Strauss & Corbin 1998) zurück und beinhaltet Frageblöcke zur Erfahrung, zum Ist-Zustand, zu einem möglichen Wunschzustand des Prozesses sowie zu ergänzenden Anmerkungen. Neben Fragen zum Prozess selbst wurde der Leitfaden um Aspekte der soziotechnischen Systemgestaltung nach Herrmann et al. Herrmann, Jahnke, and Nolte (2021) ergänzt.

Zusammenfassend bleibt festzuhalten, dass für einen Einsatz in einem realitätsnahen bis realistischen Umfeld folgende Herausforderungen zu bewältigen sind:

- Heterogene und hochdimensionale Eingabedaten
- Wenige bewertete/gelabelte Eingabedaten
- Fehlende Daten
- Adäquate UI-Gestaltung inklusive Explainable AI

Für sich betrachtet existieren zu jeder der Herausforderungen Lösungsansätze. Der Kern der Dissertation ist daher die unterschiedlichen Ansätze an die Anwendungsdomäne anzupassen, zu integrieren, sie mit einer adäquaten Benutzungsschnittstelle auszustatten, und im Anwendungsumfeld hinsichtlich Systemleistung und Usability zu erproben.

2. Untersuchungskonzept und -durchführung

Das Untersuchungskonzept folgt den eingangs erläuterten Überlegungen zur Problemstellung und wurde ausführlicher in Westhoven and Adolph (2022) dargestellt. Seit der initialen Konzeption wurde an drei größeren Bereichen gearbeitet, namentlich der Anforderungserhebung inklusive des Datenzugangs, dem algorithmenseitigen Umgang mit unstrukturiertem Text sowie dem Interaktionsdesign.

In den folgenden Unterkapiteln werden diese Bereiche noch einmal für sich beleuchtet und Zwischenergebnisse aufgezeigt.

2.1 Anforderungserhebung und Datenzugang

Zur Anforderungserhebung wurden und werden weiterhin Interviews mit Experten aus der Arbeitsschutzdomäne geführt. Dazu wurde ein Gesprächsleitfaden entworfen, der das Konzept der Grounded Theory (Strauss & Corbin 1998) aufgreift und Erfahrung, Ist-Zustand und einen möglichen Wunschzustand des Prozesses in einzelnen Blöcken abfragt. Neben Fragen zum eigentlichen Prozess werden auch Aspekte der

soziotechnischen Systemgestaltung abgefragt, wobei sich an Herrmann et al. (2021) orientiert wurde.

Es wurden eingangs u. a. aus Gründen des einfachen Zugangs Interviews mit Arbeitsschutzpraktikern des eigenen Instituts geführt. Diese sind als Arbeitsschutzpraktiker weit genug von der wissenschaftlichen Arbeit entkoppelt, um als weitgehend unbeeinflusste Informationsquelle zu dienen. Die Ergebnisse der Interviews wurden in Westhoven (2022) zusammengefasst berichtet und geben einen Überblick über den Prozess der Risikoanalysen im betrieblichen Arbeitsschutz sowie über die Anforderungen an ein Unterstützungssystem.

Vom ursprünglich vorgesehenen Kooperationsunternehmen wurde die Zusammenarbeit aus Ressourcengründen kurz vor Abschluss einer Vereinbarung für Daten- und Feldzugang beendet. Hier wurden vorab zwei Gruppeninterviews geführt, aufgrund der Kontextabhängigkeit von Risikoanalysen im betrieblichen Arbeitsschutz sind diese für sich genommen aber schlecht weiter verwertbar. Die Ergebnisse sollen daher erst zusammen mit weiteren Interviewergebnissen aus Unternehmen aufgearbeitet werden.

Derzeit laufen Gespräche mit einem multinationalen Konzern sowie einem Arbeitsschutzdienstleister, um Feld- und Datenzugang zu realisieren. Mit beiden Unternehmen wurden ebenfalls Interviews geführt, davon der Großteil als Einzelinterviews.

Aus den bisher erlangten Interviewergebnissen lässt sich bezüglich der bisherigen Arbeiten feststellen, dass die als relevant identifizierten Aspekte umfassend erschlossen wurden. Als größere Unwägbarkeit hat sich lediglich die Zeitplanung der Interviewdauer herausgestellt, da zwischen den Interviewpartnern eine hohe Varianz bezüglich der Detaillierung der Antworten bestand und der Zeitaufwand entsprechend mit variiert hat.

2.2 Umgang mit strukturiertem und unstrukturiertem Text

Ein Kernergebnis der Interviews ist durchweg, dass als Datenformat insbesondere Text vorherrscht, der mehr oder weniger strukturiert, detailliert sowie digitalisiert verfügbar ist. Es war grundsätzlich erwartet worden, dass objektive Daten vergleichsweise spärlich vorliegen, ebenso wie entsprechende Label bzw. Bewertungen dieser Daten (siehe Clemen & Winkler 1999). Tatsächlich ist das Textformat allerdings so dominant, dass sich von Algorithmenseite her ein Fokus auf das sogenannte Natural Language Processing (NLP) gebietet. NLP deckt dabei die gesamte Breite von maschineller Textverarbeitung ab und beinhaltet unter anderem Spracherkennung, Übersetzung, Information Retrieval oder auch Zusammenfassungen (Chowdhary 2020). Zuletzt wurden im NLP mit Deep Learning Fortschritte in schneller Folge gemacht (Otter et al. 2020), was nicht zuletzt in der Veröffentlichung großer Sprachmodelle wie ChatGPT (OpenAI 2023) gemündet ist.

Zum Veröffentlichungszeitpunkt der ersten großen Sprachmodelle wurden Ansätze untersucht, wie die textuelle Datenbasis mit „klassischen“ Deep Learning Methoden verarbeitet werden kann (Westhoven & Jadid 2023). Durch das Aufkommen und die seitdem schnell fortschreitende Entwicklung großer Sprachmodelle hat sich nun allerdings der Fokus dahin verschoben, ob nicht durch Transfer Learning bzw. Domain Adaption dieser Modelle bessere Ergebnisse zu erwarten sind. Einen guten Überblick zum Thema Adaption bieten Hu et al. (2021). Da sich die oben angesprochenen Probleme mit Feld- und Datenzugang ergeben haben, rückt dieser Aspekt der Dissertation erst gerade wieder in den Fokus.

Grundsätzlich bleibt es dabei, dass dieser Teil der Arbeit einen stark explorativen Charakter besitzt und sich erst in Verbindung mit den kontextbezogenen Daten aus Betrieben sagen lässt, welche unterstützenden KI-Anwendungen letztlich realisierbar sein werden. Daran hängt ebenso, wie Performance und Usability bei der Evaluation konkret gemessen werden können.

2.3 Interaktionsgestaltung

Aus den bereits geführten Interviews konnten bereits die grundsätzlichen Anforderungen an ein Interaktionsdesign für eine KI-gestützte Software abgeleitet werden. In Westhoven and Herrmann (2023) wurden diese Anforderungen mit verfügbarem Wissen zum Design von Interaktionen mit KI abgeglichen. Es wird damit aufgezeigt, auf welche Interaktionsaspekte bei der Gestaltung von KI-Nutzungsschnittstellen für das angedachte Anwendungsgebiet besonders geachtet werden muss.

Während es den Rahmen sprengen würde, die gesamte Herleitung hier wiederzugeben, lässt sich zusammenfassend festhalten, dass durch die Integration zweier Komponenten den Besonderheiten von Mensch-KI-Interaktion weitgehend Rechnung getragen wird. Zum einen wird eine sogenannte Explanation Engine benötigt, eine Programmkomponente, welche Erklärungen für das maschinelle Wirken zur Verfügung stellt. Zum anderen kann eine flexible Simulationsumgebung Explorationsmöglichkeiten im Gesamtsystem wie auch in einzelnen Fällen ermöglichen.

Wie diese beiden Programmkomponenten in der oben beschriebenen zu implementierenden NLP-Anwendung realisiert werden können, ist abhängig von der tatsächlichen Anwendung. Entsprechend ist die Weiterarbeit in diesem Bereich bis dahin zurückgestellt.

3. Methodische Überlegungen zum weiteren Verlauf

Gegenwärtig laufen mehrere Stränge parallel. Prioritär zu behandeln ist die Kooperation mit einem Kooperationsunternehmen, in dem Arbeitsschutzdaten gesammelt werden können, um deren Handhabung in KI-Systemen zur Unterstützung der Risikoanalyse zu untersuchen. Erste Gespräche haben nach Wegfall eines weit gediehenen Kontakts wieder stattgefunden und es läuft derzeit eine Prüfung, welche vom Dienstleister betreuten Unternehmen bzw. welche Bereiche des Großunternehmens infrage kommen und welche Daten überhaupt zur Verfügung gestellt werden können. Es wurden hier bereits weitere Experteninterviews geführt, um die Perspektive von Arbeitsschutzpraktikern in Unternehmen ergänzen zu können. Eine Zusammenfassung der Ergebnisse steht dabei noch aus.

Die Entwicklung sinnvoller und lauffähiger KI-Systeme zur Unterstützung von Risikoanalyseprozessen im betrieblichen Arbeitsschutz gelangt mit den fortschreitenden Gesprächen mit den Unternehmen und nach der Umfokussierung auf die Möglichkeiten von großen Sprachmodellen wieder eine höhere Priorität. Diese zu implementierenden Systeme werden im Verlauf des Projektes auch mit dem Unternehmen vor Ort evaluiert, um weitere Einblicke in Potenziale, aber auch Fallstricke von KI in diesem Bereich zu gewinnen.

In Ergänzung zu den aufgezeigten Bereichen der Dissertation ist durch den kooperierenden Arbeitsschutzdienstleister bereits ein belastbarer Kontakt zu Arbeitsschutzexperten auf KMU-Ebene ist vorhanden. Um Aussagen über die Generalisierbarkeit

der Anwendung zu erlangen, soll hier noch zusätzlich erhoben werden, ob Methoden auch auf in KMUs realistischerweise vorzufindenden / zu erhebenden Daten lauffähig sind und falls nein, ob man sich algorithmisch oder auch betrieblich entsprechend anpassen kann.

4. Literatur

Chowdhary K (2020). Natural language processing. Fundamentals of artificial intelligence, 603–649.
Clemen RT & Winkler RL (1999). Combining probability distributions from experts in risk analysis. Risk analysis, 19(2), 187–203.
Dellermann D, Ebel P, Söllner M & Leimeister JM (2019). Hybrid intelligence. Business & Information Systems Engineering, 61(5), 637–643.
Doran D, Schulz S & Besold TR (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv: 1710.00794.
Herrmann T, Jahnke I & Nolte A (2021). A problem-based approach to the advancement of heuristics for socio-technical evaluation. Behaviour & Information Technology, 1–23.
Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, … Chen W (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv: 2106.09685.
Kingma DP & Welling M (2013). Auto-encoding variational bayes. arXiv preprint arXiv: 1312.6114.
OpenAI (2023). ChatGPT (Version 4). Retrieved from https://openai.com
Otter DW, Medina JR & Kalita JK (2020). A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, 32(2), 604–624.
Schmidt A & Herrmann T (2017). Intervention user interfaces: a new interaction paradigm for automated systems. Interactions, 24(5), 40–45.
Shams R (2014). Semi-supervised classification for natural language processing. arXiv preprint arXiv: 1409.7612.
Shneiderman B, Plaisant C, Cohen MS, Jacobs S, Elmqvist N & Diakopoulos N (2016). Designing the user interface: strategies for effective human-computer interaction: Pearson.
Strauss A & Corbin J (1998). Basics of qualitative research techniques.
Wang Y, Yao Q, Kwok JT & Ni LM (2020). Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 53(3), 1–34.
Westhoven M (2022). Requirements for AI Support in Occupational Safety Risk Analysis. In Proceedings of Mensch und Computer 2022 (pp. 561–565).
Westhoven M & Adolph L (2022). Concept for Supporting Occupational Safety Risk Analysis with a Machine Learning Tool, Cham.
Westhoven M & Herrmann T (2023). Interaction design for hybrid intelligence: the case of work place risk assessment. Paper presented at the International Conference on Human-Computer Interaction.
Westhoven M & Jadid A (2023). Supporting Work Place Risk Assessments by Means of Natural Language Processing. Paper presented at the 69th GfA Frühjahrskongress, Hannover, Germany.

Gesellschaft für Arbeitswissenschaft e.V.

Arbeitswissenschaft in-the-loop

Mensch-Technologie-Integration und ihre Auswirkung auf Mensch, Arbeit und Arbeitsgestaltung

70. Kongress der
Gesellschaft für Arbeitswissenschaft e.V.

Institut für Arbeitswissenschaft und
Technologiemanagement IAT
Universität Stuttgart

In Zusammenarbeit mit dem Fraunhofer-Institut für
Arbeitswirtschaft und Organisation IAO

06. – 08. März 2024

GfA-Press

Bericht zum 70. Arbeitswissenschaftlichen Kongress vom 06. – 08. März 2024

Institut für Arbeitswissenschaft und Technologiemanagement (IAT), Universität Stuttgart
In Zusammenarbeit mit: Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO), Stuttgart

Herausgegeben von der Gesellschaft für Arbeitswissenschaft e.V.
Sankt Augustin: GfA-Press, 2024
ISBN 978-3-936804-34-8

NE: Gesellschaft für Arbeitswissenschaft: Jahresdokumentation

Als Manuskript zusammengestellt. Diese Jahresdokumentation ist nur in der Geschäftsstelle (s. u.) erhältlich.

Geschäftsstelle der GfA
Simone John, Tel.: +49 (0)30 1300-13003, Alte Heerstraße 111, D-53757 Sankt Augustin
info@gesellschaft-fuer-arbeitswissenschaft.de · www.gesellschaft-fuer-arbeitswissenschaft.de

Screen design und Umsetzung
© 2024 fröse multimedia, Frank Fröse,
office@internetkundenservice.de, www.internetkundenservice.de