Organizational Safety and Health Topics in Current German Artificial Intelligence Projects and the Road Ahead

Martin WESTHOVEN, Silvia VOCK, Lars ADOLPH

Federal Institute for Occupational Safety and Health,
Friedrich-Henkel-Weg 1-25, 44149 Dortmund, Germany

Abstract. As Artificial Intelligence spreads into more and more areas of our lives, not least our work lives, political and regulatory aspects are in need of revision. We analyzed the current German AI project landscape to assess if and how safety and health aspects are considered today. Finding only a small amount of projects, which also address limited aspects of safety and health, we compiled theoretical considerations on safety-relevant characteristics of Artificial Intelligence from literature to further map out possible future research areas.

Keywords: Occupational Safety and Health, Artificial Intelligence, Project Landscape Analysis

1. Introduction

Advances in the field of Artificial Intelligence (AI) generate pressure to apply the technology in more and more areas. As such, numerous political and regulatory aspects are in need of revision, highlighted by efforts such as the European Commission's White Paper on Artificial Intelligence or the national AI-Strategy of the German Federal Government. One such aspect is the appropriate consideration of occupational safety and health (OSH) when deploying of AI. A key point of discussion is the assessment of risks associated with various AI functionalities. The characterization of relevant features is the basis for decisions about a conformity assessment process that may be required for the European market. In principle, three levels of assessment are possible: a.) First party: self-assessment, b.) Second party: review by the customer or user, and c.) Third party: review by independent third parties. Adequate criteria and processes are important for the process of ensuring product safety in Europe. Currently, however, it is difficult to assess in which areas of application particularly risky AI applications are to be found at all. Thus, we have investigated the German AI research landscape with a focus on risks regarding OSH. We provide an overview over the current research activities in Germany and add theoretical considerations to map out future research topics.

2. Organizational Safety and Health in Germany's Current Project Landscape

We analyzed 600 AI projects compiled by the "Plattform Lernende Systeme" according to their relevance for safety and health in an occupational context. Their relevance was assessed by answering the following questions:

- Does the project cover AI-applications, from which potential dangers can arise?

- Does the project cover AI-applications which have the objective or potential to be used to avoid hazards? Here we distinguish between product safety and operational safety.
- Does the project cover AI properties of general nature, which are not connected to a specific application context? Projects, which address e.g. the performance of AI, transparency, explainability and dependability fall into this category.

A schematic overview of these categories is given in Figure 1. During evaluation the projects were assigned to one of these categories and labeled either “relevant” or “indirectly” relevant. Projects which did not fit into one of these categories were labelled “irrelevant”. A relevant project must address safety directly (independent from the actual application) or AI has to be used as a safety function. Due to their potentially direct impact on safety functions (Kasper & Voß, 2018), projects on security are included in this survey as well. Furthermore, projects addressing clearly safety-critical applications are likewise labelled relevant. Indirectly relevant are projects, which address AI-applications in e.g. industrial context with potential for integrating safety aspects but without discussing this topic explicitly. Furthermore, an indirectly relevant project addresses innovative AI-methods from a “non-operational” context, but with potential to avoid hazards, e.g. when using data with high dimensionality, complexity, or heterogeneity or methods are suggested, which are inherently suitable e.g. for anomaly detection.

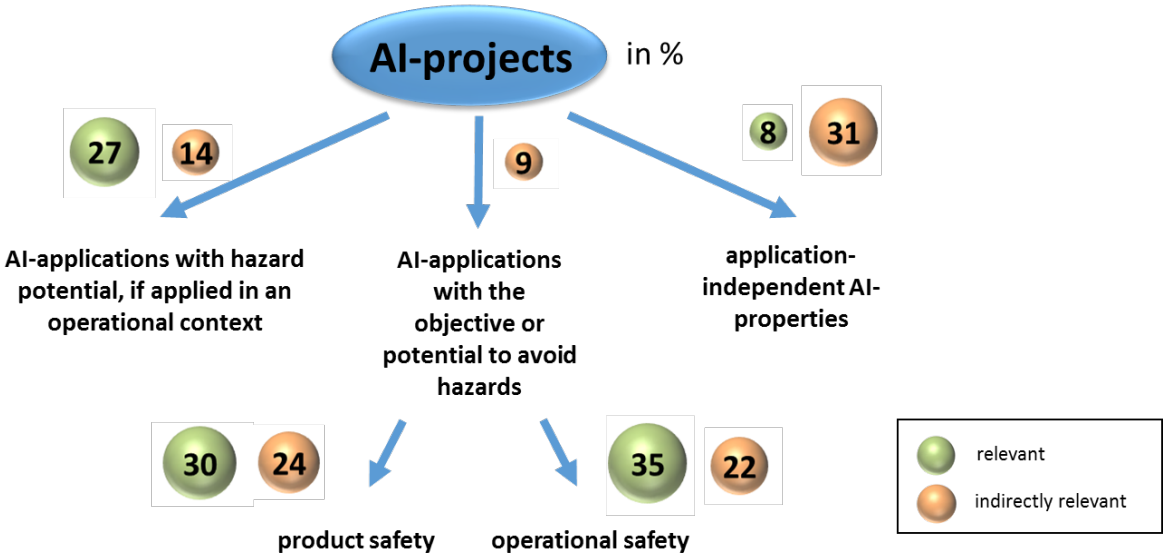

Figure 1. *Schematic overview of the pre-defined categories and the respectively assigned projects given relative to the total number in their category (bubbles above the categories, 169 indirectly and 37 relevant projects in total).*

In addition to the rating of the projects, key words were assigned to the relevant or indirectly relevant projects, which describe their main contribution area. The evaluation was performed as an expert review. The quality of agreement between the reviewers was measured by calculating the interrater reliability. The weighted Cohen’s Kappa resulted in 0.422, which is an indicator for moderate agreement (Landis & Koch, 1977).

Out of 600 AI projects, 169 were rated indirectly relevant and 37 are seen as relevant. Figure 1 gives an overview of the project distribution with regard to the pre-defined categories and to the rating (relevant, indirectly relevant). The projects labelled as relevant are almost equally distributed between the categories “AI application with

hazard potential", "product safety" and "operational safety". The indirectly relevant projects are also quite equally distributed between 3 categories, but here the category "AI applications with hazard potential" is replaced by the category "AI properties". This is obviously related to the nature of indirectly relevant project goals, which are more likely to be application independent.

With regard to the large potential of AI in an occupational context, it is remarkable that only 6 % of the assessed projects in the German research landscape show a direct relation this topic. Furthermore, one third of this already small number are projects with hazard potential when applied in an operational context. However, almost one third of the projects were rated indirectly relevant to the occupational context and therefore are supposed to have potential to be applied in a wider range in future than actually addressed by the researchers.

The current focus of German research activities with relation to occupational safety and health topics is visualized by assigning all the relevant and indirectly relevant projects together with their respective keywords to the pre-defined categories (see Fig. 2).

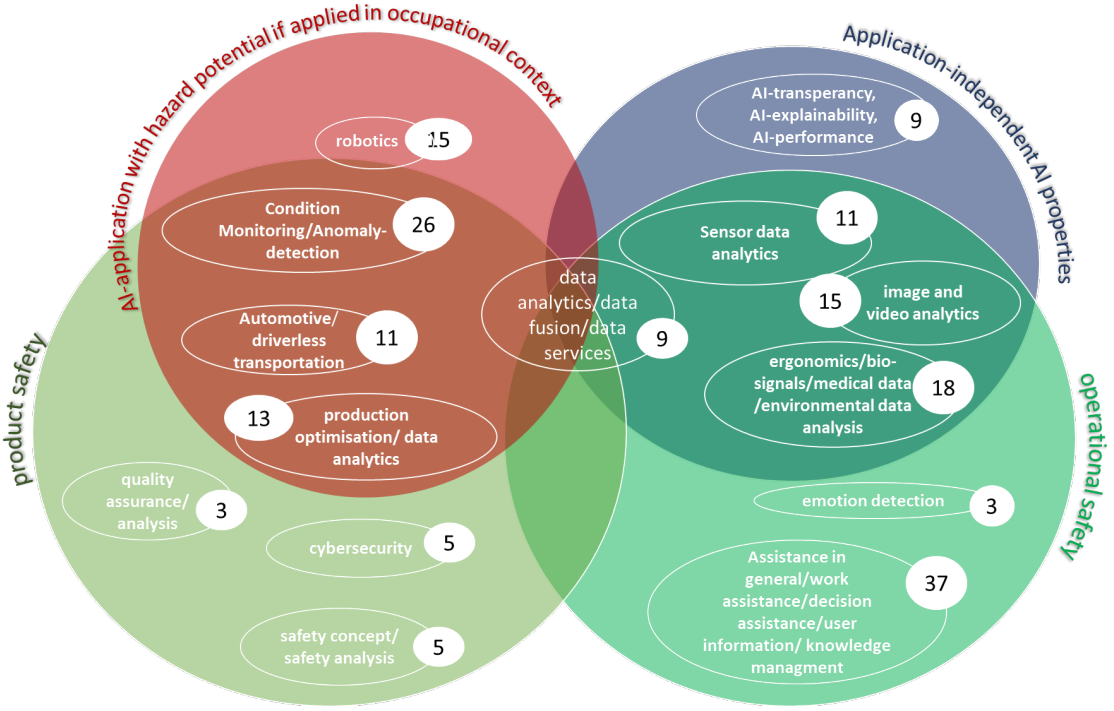

Figure 2. *Schematic overview of the German research landscape (relevant and indirectly relevant projects) with relation to occupational safety and health topics.* 22 projects were omitted in this representation, because they were too special and did not fit into this generalized view.

Having identified this apparent lack of research focusing the safety of AI itself, we compiled further theoretical considerations from literature to guide future research to a more comprehensive view of safe AI. We roughly distinguish technology-inherent characteristics of AI from those dependent on context. While the former mostly influence error probability, the latter mostly affect the damage potential.

With relevance to product safety, the field of condition monitoring and anomaly detection (26 projects) has the largest share, while in the field of operational safety the AI-algorithms with assistance functionalities (37 projects) are most represented. Very few robotics projects address safety in any form explicitly.

Again it is remarkable that almost one third of the projects show a hazard potential if applied in an operational context, without explicitly addressing it. Furthermore, the overview visualizes the emphasis on operational safety and health topics, while product safety and more precisely safety concepts, safety analysis methods, AI dependability and cybersecurity aspects are only rarely in the focus of today's research activities.

3. Safety Relevant Characteristics of Artificial Intelligence

Having identified this apparent lack of research focusing the safety of AI itself, we compiled further theoretical considerations from literature to guide future research to a more comprehensive view of safe AI. We roughly distinguish technology-inherent characteristics of AI from those dependent on context. While the former mostly influence error probability, the latter mostly affect the damage potential.

3.1 Technology-Inherent Characteristics

Where human autonomy means setting goals for oneself, autonomy for AI typically refers to degrees of freedom in deciding how to reach a goal (Totschnig, 2020). Following Franklin and Graesser (1996), the most simple case is the binary choice of executing an action or not. More complex variants could also allow to alter parameters of an action or even to devise entirely new actions. The safety relevance of this characteristic stems from the problem of posing a well-defined goal including desirable constraints (see e.g. Amodei et al. (2016)). The capability of a program to adapt itself to compensate for varying operational conditions is called adaptivity and is related to autonomy. Adaptivity can range from using pre-defined parameter sets to the extreme case of being able to fully alter the program code. Safety is affected, if the effects of the adaption can't be reliably or efficiently predicted, as manifest in large parameter spaces.

Temporal continuity of software programs can allow for error propagation to evolve for longer intervals and is thus a relevant safety characteristic. While temporal continuity is defined as a base requirement for software agents by Franklin and Graesser (1996), AI systems don't necessarily need full temporal continuity. A trained system wouldn't need continuity between its inference-time activations.

Interaction of an AI system can range from forms of training to runtime-interaction with other systems, including humans. A thorough taxonomy in this regard is provided by Dellermann et al. (2019). Interaction is to be regarded as a source of outside or human error, as well as it is an attack vector. Interaction can also influence human users in their use of the system. Safe operation of a system can thus indirectly be affected by itself.

Transparency influences safety by allowing for an intuitive understanding of the system's operations. According to Ventocilla et al. (2018) transparency should distinguish transparency for the end-user and expert interpretability. Interpretability then means that experts can understand the algorithm as well as its decisions. If an AI system is not interpretable, its safeness can't be assessed. A special case is the verification of AI systems, as the designer formally proves the correct operation in regards to its specification.

Machine learning drawing on multiple modalities for input is called multimodal machine learning. It can mitigate weaknesses of single modalities, but also introduce new

sources for faults at the transfer between and the fusion of different modalities. Baltrušaitis, Ahuja, and Morency (2018) provide a taxonomy to classify aspects of multimodality in machine learning.

After training, weaknesses leading to faulty classifications can remain in a learned solution. Such weaknesses can lead to unintended system behavior, either unintentionally or by being exploited (see Tabassi, Burns, Hadjimichael, Molina-Markham, and Sexton (2019)). Different approaches exist to avoid such weaknesses, which are systematically studied by e.g. Pitropakis, Panaousis, Giannetsos, Anastasiadis, and Loukas (2019) and by Tabassi et al. (2019). The implementation of countermeasures generally reduces system performance.

3.2 Context-Dependent Characteristics

Data poisoning, the introduction of erroneous samples during training, can result in a faulty model (Tabassi et al., 2019). This can be avoided by access restrictions, but these can be difficult to realize for publicly governed data. An approach to detect poisoning is to check for suspiciously high error rates introduced by single data points (Pitropakis et al., 2019). Access restrictions govern the possibility of externally introduced change to the system, as e.g. in over-the-air-updates. Lacking situational awareness of external actors and the additional attack vector can affect safety.

The operational environment of AI systems can be very diverse and influences, if, and how much damage results from a system failure and who or what is affected. A general distinction can be made between physical and virtual environments. Where AI systems perform physical actions, they typically generate higher risk. Nevertheless, also virtual actions, such as stock trading, can have severe consequences. Tabassi et al. (2019) describe four types of consequences of attacks on machine learning: Integrity, availability, confidentiality, and privacy. The environment determines which problems can come into effect and how severe the consequences can become. It should be noted, that dynamic operational environments pose a significant challenge in this regard. Another aspect of the operational environment is social context, which can be safety relevant as well. It encompasses who is affected by the system and how this influences human-machine and inter-human interaction.

4. Discussion and Summary

It becomes clear that identification of specific applications, design features and functionalities of AI is necessary in order to identify criteria for assessing risks and criticality. Threshold levels for the criticality evaluation or risk assessment should be discussed with the aim to choose adequate processes for market access. This requires an analysis taking into account the complexity of determining such thresholds in different application contexts. Our project landscape analysis highlights current work in Germany towards AI as a safe-to-use tool. We complement these with considerations drawn from literature regarding the safety-relevance of AI systems to provide a comprehensive overview over current research and possible future research directions at the intersection of AI and OSH.

5. References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:.06565.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis machine intelligence, 41(2), 423-443.

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2019). The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems.

Franklin, S., & Graesser, A. (1996). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. Paper presented at the International Workshop on Agent Theories, Architectures, and Languages.

Kasper, B., & Voß, S. (2018). Neue Anforderungen an die Sicherheitsnachweisführung von Maschinen und Anlagen im Kontext von Industrie 4.0. Sicher ist Sicher, 09.2018, 368.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159-174.

Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E., & Loukas, G. (2019). A taxonomy and survey of attacks against machine learning. Computer Science Review, 34, 100199.

Tabassi, E., Burns, K. J., Hadjimichael, M., Molina-Markham, A. D., & Sexton, J. T. (2019). A Taxonomy and Terminology of Adversarial Machine Learning. In: NIST IR.

Totschnig, W. (2020). Fully autonomous AI. Science and Engineering Ethics, 26(5), 2473-2485.

Ventocilla, E., Helldin, T., Riveiro, M., Bae, J., Boeva, V., Falkman, G., & Lavesson, N. (2018). Towards a taxonomy for interpretable and interactive machine learning. Paper presented at the XAI Workshop on Explainable Artificial Intelligence.

Acknowledgements. This work is part of the “Technical and Organizational Occupational Safety” chapter of the priority program “Occupational Safety and Health in a Digitized World of Work” at the Federal Institute for Occupational Safety and Health. We thank our colleagues Swantje Robelski, Sabine Sommer, and Stefan Voß for their valuable input.

Gesellschaft für
Arbeitswissenschaft e.V.

Arbeit HUMAINE gestalten

67. Kongress der
Gesellschaft für Arbeitswissenschaft

Lehrstuhl Wirtschaftspsychologie (WiPs)
Ruhr-Universität Bochum

Institut für Arbeitswissenschaft (IAW)
Ruhr-Universität Bochum

3. - 5. März 2021

GfA-Press

Bericht zum 67. Arbeitswissenschaftlichen Kongress vom 3. - 5. März 2021

Lehrstuhl Wirtschaftspsychologie, Ruhr-Universität Bochum
Institut für Arbeitswissenschaft, Ruhr-Universität Bochum

Herausgegeben von der Gesellschaft für Arbeitswissenschaft e.V.
Dortmund: GfA-Press, 2021
ISBN 978-3-936804-29-4

NE: Gesellschaft für Arbeitswissenschaft: Jahresdokumentation

Als Manuskript zusammengestellt. Diese Jahresdokumentation ist nur in der Geschäftsstelle erhältlich.

Schriftleitung: Matthias Jäger
im Auftrag der Gesellschaft für Arbeitswissenschaft e.V.

Screen design und Umsetzung
© 2021 fröse multimedia, Frank Fröse
office@internetkundenservice.de · www.internetkundenservice.de