Mind the Economic Safety Gap

Accelerating Safety Innovation using Generative Artificial Intelligence

Andreas Kreutz, Christian Drabek, René Beck
Fraunhofer Institute for Cognitive Systems IKS
Munich, Germany
{*firstname.lastname*}@iks.fraunhofer.de

***Abstract*—The accelerating pace of technological innovation – driven in large parts by artificial intelligence – means that safety engineering is becoming a substantial cost factor in many development projects. Addressing this issue will require contributions from safety engineering research to ensure that new technologies can be safely used in safety-critical systems. In this paper, we describe the impact that a lack of novel safety engineering methods would have on products in the future. Additionally, we propose using generative artificial intelligence to create an assistant for safety engineers, which would be an effective measure to address the issue of safety assurance for complex systems. Finally, we explain remaining research challenges and how they could be addressed.**

Keywords*—*safety engineering, generative artificial intelligence

I. Introduction

In today's rapidly evolving technological landscape, the pace of innovation in fields such as mobility, automation, and health is unprecedented. The introduction of cognitive systems, AI, and machine learning has pushed the boundaries of what is possible. Companies want to capitalize on the potential economic gains of this technology, however, for safety-critical embedded systems, AI-based and AI-generated functions raise major safety concerns.

Historically, safety innovation has always been a step behind technology advancements. This creates a problem for companies – the "Economic Safety Gap". This gap represents the economic losses that companies incur when safety concerns prevent them from applying new technologies. In the past, this gap was manageable, however, due to the rapidly accelerating pace of innovation, the economic safety gap is widening, as safety research is struggling to keep up.

In this paper, we explain what impact this might have on the development of safety-critical systems and describe a way to bridge the gap: Using generative artificial intelligence (genAI) to automate parts of the safety engineering workflow. We present the topics that we are currently working on at the Fraunhofer Institute for Cognitive Systems IKS and point out the remaining research and development challenges.

II. The Economic Safety Gap

In this section, we explain the concept of the economic safety gap in more detail using an example from the robotics domain. Then, we describe ways to address the gap.

A. Overview

The economic safety gap is the difference between the systems that can be built using current technology and the ones that are provably safe to operate. Figure 1 shows a visualization of the concept and how it is developing over time. Technological innovation (green) makes it possible to develop more and more capable systems. In recent years, the pace of technological innovation has greatly accelerated due to hardware improvements, novel programming paradigms, and artificial intelligence functions. At the same time, innovation in the field of safety engineering (blue) has hardly made much progress in comparison. This creates the economic safety gap (orange): Though more productive systems are technologically feasible; they cannot be deployed because the safety of the systems cannot be assured. The gap represents the economic loss that companies incur because they cannot use novel technologies in their products. While in the past (point A in the figure), the gap was acceptably small, today (point B) and especially based on future projections (point C), safety is becoming a major blocker for innovation.

B. Example

The economic safety gap can already be observed in many domains, such as in robotic automation. Industrial robots are a well-established technology for automating dull, dirty, and dangerous tasks in manufacturing, such as spot welding in automotive manufacturing. Recently, companies are exploring robotic automation in many other applications, ranging from palletizing to machine tending and assembly. These use cases have demanding requirements, such as high flexibility or low retooling effort. Space restrictions and the need for human-robot

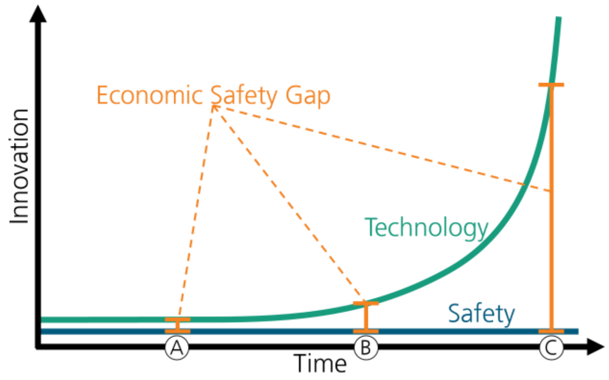


Fig 1: The economic safety gap.

collaboration require new, complex safety concepts that enable fenceless operation at high speeds.

These use cases are all technically feasible due to advancements in robot hardware, highly accurate sensors, control abstraction layers enabling low-/no-code programming, and by using AI functions for perception and control. However, these capabilities also greatly increase the complexity of the application, which oftentimes means that established safety standards do not apply. Instead, a customized risk assessment, safety concept derivation, and verification and validation becomes necessary, which requires significant effort. In many cases, only large companies can manage the costs associated with this effort. As a result, especially small and medium-sized enterprises (SMEs) are unable to capitalize on the potential productivity gains through robotic automation.

C. *Addressing the gap*

The example illustrates that the economic safety gap is already a problem today, which will only worsen based on projections of the future (see Figure 1). There are three ways how the gap might be addressed.

If we *slow down technology innovation*, we acknowledge that safety is a necessary requirement for system deployment which must not be disregarded for economic gains. However, this approach comes with a large amount of unused technological potential, resulting in significant economic losses. Additionally, due to the proportionally larger impact of safety engineering on the overhead costs of SMEs, we might witness reduced competitiveness of SMEs in the global economy.

Therefore, it seems likely that *accepting the growing safety vacuum* might become an attractive solution. To remain competitive, companies might decide to deploy systems with less comprehensive safety concepts than those of conventional systems. In this scenario, technology innovators decide what is “safe enough” for their applications, at the detriment of system operators. Though this might be economically preferable, it will hardly be possible in highly regulated markets such as the European Union [1].

Finally, we could *push for innovation in safety engineering* to not only innovate on what is technologically feasible, but also what is provably safe to use. Within safety engineering research, there are several active directions, such as dynamic risk management [2], [3], model-based safety engineering [4], [5], and safety of machine learning [6], [7]. In addition to these contributions to functional safety, innovation is also required with regards to making the engineering process more efficient.

In summary, the only acceptable approach to closing the economic safety gap is accelerating safety engineering. This ensures that technological progress can effectively be used to create economic gains without compromising safety. Achieving this will require many contributions. In the next section, we present one approach that aims to automate parts of the engineering process using genAI.

III. Automating Safety Engineering with Generative AI

GenAI is a subfield of artificial intelligence which uses generative models to produce text, images, videos, or other forms of data. Especially Large Language Models (LLMs) have been able to achieve human-like performance in natural language generation because of novel model architectures that have enabled training on very large amounts of data. Ever since the release of ChatGPT [8], such models have also received a lot of attention from industry for creating AI assistants for numerous tasks, such as marketing, software development and customer support [9]. Similarly, genAI has a lot of potential for safety engineering.

A. *Safety Engineering Workflow*

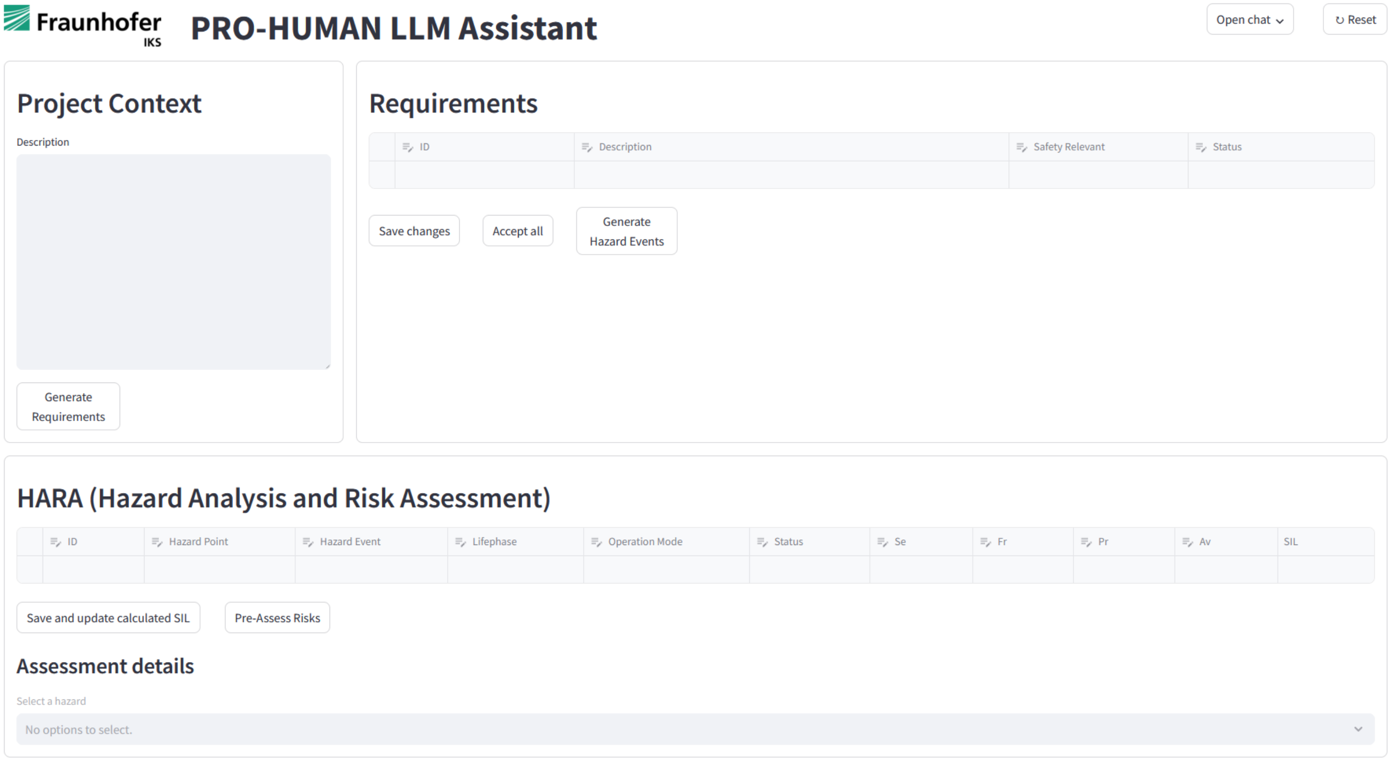
When developing a safety-critical system, special care must be placed on mitigating risks to its environment. IEC 61508 [10] recommends several activities within a comprehensive safety lifecycle to achieve a tolerable level of residual risk through the application of E/E/PE safety-related systems. This engineering workflow results in a safety case that documents all steps and provides convincing evidence that a system is safe for use. In this paper, we focus on the early stages of the safety lifecycle.

After defining the scope of the system, the first step is a detailed hazard and risk analysis (HARA) to identify potential hazardous events and assess the associated risks with categories such as severity of consequences and frequency of exposure. Based on this classification, a Safety Integrity Level (SIL) is derived as a necessary requirement for the function mitigating each hazard. IEC 61508 recommends that the HARA is performed in workshops with a variety of stakeholders, e.g., developers, operators, and occupational safety representatives. This means that this activity takes a lot of time and effort, and its result is very dependent on the expertise of the participants. Additionally, it is greatly impacted by the economic safety gap: More complex systems with advanced capabilities widen the scope of the required HARA substantially and, thus, greatly complicate the analysis.

B. *Assistant Prototype*

We propose that a genAI-based assistant could alleviate these issues and support safety engineers in performing HARAs. To this end, at Fraunhofer IKS, we have developed a proof-of-concept implementation of such a tool, shown in Figure 2. The

This work was funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Institute for Cognitive Systems.


PRO-HUMAN LLM Assistant

Project Context

Description

Requirements

ID	Description	Safety Relevant	Status

HARA (Hazard Analysis and Risk Assessment)

ID	Hazard Point	Hazard Event	Lifephase	Operation Mode	Status	Se	Fr	Pr	Av	SIL

Assessment details

Select a hazard

No options to select.

Fig 2: GenAI-based safety engineering assistant.

tool demonstrates how a workflow supported by genAI could look like. Safety engineers can use it to collaboratively work with an LLM in a similar way to how a HARA workshop would be conducted. The tool converts the interactions of the engineer into appropriate prompts to interface with the LLM as a backend. Figure 3 shows an example of such an interaction. In this interaction, the use case description and requirements provided by the engineer in the top left and right boxes of Figure 2 have been compiled into textual context that is provided to the LLM for analysis. A predefined prompt and prompt preamble is used to retrieve information from the model. The tool also features a chat function with which the user can ask questions about the output of the assistant and request corrections.

C. *Potential Benefits*

We are currently in the process of evaluating the usefulness of the tool with interested partners from industry. Based on our own experience and feedback, the tool could bring several benefits: Automating parts of the engineering workflow can streamline processes, *save resources* and potentially reduce the time needed for safety assurance from months to weeks. At the same time, this efficiency gain also enables safety assurance for *more complex applications* than previously possible. HARAs that would previously have been too expensive to perform could become economically viable with the support of genAI. Furthermore, the tool could *strengthen the safety case* because using genAI could result in a more thorough exploration of the analysis space, especially if the genAI agent is provided with data of previous analyses. Finally, this would also *reduce the reliance on individual expertise*, which would enable more companies to access rigorous and reliable safety analyses.

IV. Research Challenges

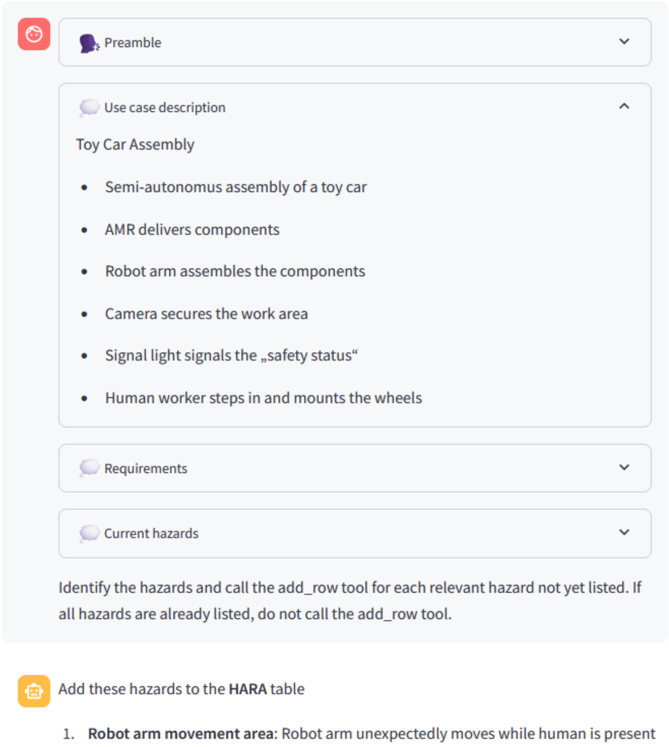
Using current LLMs as a backend for the safety assistant already enables the tool to give useful support for engineers and competently answer questions about its output. However, for a productive use in industrial applications, further development will be required. Additionally, there are several research challenges that need to be tackled to make the results more reliable.

A. *Improve Trustworthiness*

General LLMs are currently limited by the reliability of their output. Such models tend to hallucinate [11], which means that they generate information that is false, misleading, or nonsensical, despite it being presented in a coherent manner. User of LLMs, especially in safety-critical applications, expect output that can be trusted to be correct. An important line of research therefore is concerned with increasing the trustworthiness of the output of genAI. Application-specific engineering of LLM tools, such as prompt engineering or Retrieval Augmented Generation, can increase the pool of information that is accessible to the LLM, which can reduce the tendency to hallucinate. Additionally, approaches such as chain-of-thought or attribution force the agent to provide reasoning and sources for its output, which makes it easier for the user to verify the correctness of the output. First results indicate that such techniques successfully contribute to improving the reliability of the LLM output [12].

B. *Facilitate Collaboration*

As genAI-based tools cannot provide guarantees for their output, the workflow for using such tools needs to be designed



Preamble

Use case description

Toy Car Assembly

- Semi-autonomus assembly of a toy car
- AMR delivers components
- Robot arm assembles the components
- Camera secures the work area
- Signal light signals the „safety status“
- Human worker steps in and mounts the wheels

Requirements

Current hazards

Identify the hazards and call the add_row tool for each relevant hazard not yet listed. If all hazards are already listed, do not call the add_row tool.

Add these hazards to the **HARA** table

1. **Robot arm movement area**: Robot arm unexpectedly moves while human is present

Fig. 3: Example for an interaction with the LLM. The prompt is composed by the tool based on the input from the user.

in a way that the analysis results are always verified by a human safety engineer. An overreliance on the tool must be avoided by keeping humans engaged in the engineering process. Instead of generating a complete solution, an iterative approach is more sensible: Once an initial solution is auto-generated, human feedback is vital to further improve it and align the result with what the user intended. This also alleviates acceptance issues, as engineers need not worry that they will be replaced by AI. In addition, training and education can enable humans to use AI tools most effectively. In the European Union, this has recently become a mandatory measure according to the EU AI act [13]. In terms of interface design, collaboration needs to be facilitated by integrating the genAI assistant into the existing workflow of users.

C. *Build Specialized Models*

Using pretrained LLMs for safety engineering tasks already works well, because these models are able to replicate human-like reasoning on natural language generation tasks. However, application-specific knowledge can be difficult to access, even with retrieval augmentation generation methods. The logical next step, therefore, is to train a safety-engineering-specific model by fine-tuning the pretrained weights of a generic LLM. For this, a sufficiently large and diverse set of training data is necessary. Data from previous safety analyses is oftentimes confidential and companies do not wish to share such intellectual property with competitors, which complicates this approach. Datasets from individual companies will hardly be sufficient to fine-tune a large model like an LLM.

However, as products become more and more standardized, functional safety becomes a basic requirement instead of a differentiating factor, as proven by the numerous safety standards available in all domains. Similarly, a standardized safety assistant could become an enabling technology for companies to develop more complex products that differentiate themselves based on their features.

V. Conclusion

Safety engineering continues to be an important aspect of product development. Novel technologies revolutionize what is technologically feasible, however, also push established safety engineering methods to their limits. In this paper, we have described the economic consequences of the technological developments we are witnessing. Furthermore, we have proposed a genAI-based safety assistant as an approach to extend the limitations of established safety methods. Our prototypical tool can be used to support the work of safety engineers, saving resources, enabling more complex applications, strengthening safety cases, and using expertise that is rare to come by more effectively.

Several research challenges remain for a productive use. At Fraunhofer IKS, we are actively investigating how to make genAI more reliable to use, paving the way for its adoption in safety-critical applications.

References

[1] European Parliament and the Council, “Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on Machinery and Repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC,” European Union, Jun. 2023.

[2] G. Weiss, P. Schleiss, D. Schneider, and M. Trapp, “Towards Integrating Undependable Self-Adaptive Systems in Safety-Critical Environments,” in Proceedings of the 13th International Conference on Software Engineering for Adaptive and Self-Managing Systems, Gothenburg Sweden: ACM, May 2018, pp. 26–32. doi: 10.1145/3194133.3194157.

[3] M. Trapp and G. Weiss, “Towards Dynamic Safety Management for Autonomous Systems,” Saf.-Crit. Syst. Club, 2019.

[4] A. Kreutz, G. Weiss, and M. Trapp, “Automatic Deduction of the Impact of Context Variability on System Safety Goals,” in 2024 19th European Dependable Computing Conference (EDCC), Leuven, Belgium: IEEE, Apr. 2024, pp. 1–8. doi: 10.1109/EDCC61798.2024.00015.

[5] G. Weiss, M. Zeller, H. Schoenhaar, C. Drabek, and A. Kreutz, “Approach for Argumenting Safety on Basis of an Operational Design Domain,” in Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI, Lisbon Portugal: ACM, Apr. 2024, pp. 184–193. doi: 10.1145/3644815.3644944.

[6] A. Salvi, G. Weiss, and M. Trapp, “Adaptively Managing Reliability of Machine Learning Perception under Changing Operating Conditions,” in 2023 IEEE/ACM 18th Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS), Melbourne, Australia: IEEE, May 2023, pp. 79–85. doi: 10.1109/SEAMS59076.2023.00019.

[7] T. Haider, K. Roscher, B. Herd, F. Schmoeller Roza, and S. Burton, “Can you trust your Agent? The Effect of Out-of-Distribution Detection on the Safety of Reinforcement Learning Systems,” in Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, Avila Spain: ACM, Apr. 2024, pp. 1569–1578. doi: 10.1145/3605098.3635931.

[8] OpenAI, “Introducing ChatGPT.” Accessed: Feb. 19, 2025. [Online]. Available: openai.com/index/chatgpt/

[9] D. Maliuagina, “45 real-world LLM applications and use cases from top companies.” Accessed: Feb. 19, 2025. [Online]. Available: www.evidentlyai.com/blog/llm-applications

[10] DIN IEC 61508, “Functional Safety of Electrical/Electronic/ Programmable Electronic Safety-Related Systems,” Deutsches Institut für Normung, 2006.

[11] Z. Ji et al., “Survey of Hallucination in Natural Language Generation,” ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[12] B. Balu et al., “Towards Automated Safety Requirements Derivation using Agent-Based RAG,” accepted for publication at AAAI-MAKE, 2025.

[13] European Parliament and the Council, “Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act),” European Union, Jun. 2024.