Chapter 6
Artificial Intelligence (AI)

This chapter first provides a section on artificial intelligence (AI) in high risk systems, giving an overview over the current progress in standards relating to this topic. Next, a section addresses explainable AI (XAI) both as a technical concept and as a concept that has evident human and organizational sides to it. Lastly, a section on the concept of safety of intended functionality (SOTIF) is provided as it addresses safety in AI-driven systems, especially autonomous vehicles. This approach helps mitigate risks from functional insufficiencies in AI algorithms, making it vital for deploying AI safely in high risk areas.

6.1 Safe Artificial Intelligence (AI)

Objective
The objective of this part of the safety plan is to provide guidelines for the safe integration of AI technologies in safety-related systems. It aims to foster awareness of the properties, functional safety risk factors, available functional safety methods, and potential constraints associated with AI technologies. The plan should outline a structured approach to ensure that AI systems comply with existing functional safety standards and address the unique challenges posed by AI, including machine learning. By implementing this plan, developers can mitigate risks and ensure the reliability and safety of AI-driven safety functions.

Information
This chapter has a direct link to the safety case (SC) chapter with the same title. Other relevant documents depend on the product or system to be developed and, e.g., the classes presented in ISO/IEC TR 5469:2024.

It is worth noting that the ISO/PAS 8800:2024-12 Road Vehicles, "Safety and artificial intelligence," has recently been issued. This pas includes many work products that significantly impact the field of AI safety for road vehicles.

T. Myklebust et al., *The AI Act and The Agile Safety Plan*, SpringerBriefs in Computer Science, https://doi.org/10.1007/978-3-031-80504-2_6

‘AI system’ means, according to the AI Act, *a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit* ***adaptiveness after deployment,*** *and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*.

Although adaptiveness after deployment is an important part of the “AI system” definition, the EU AI Act does not impose specific requirements directly tied to “adaptiveness after deployment” as a stand-alone feature. However, the concept of adaptiveness is integral to the definition of AI systems in the Act, which states that AI systems may exhibit adaptive behaviours after deployment. General requirements are related to risk management and to ensuring that AI systems consistently meet safety, transparency, accuracy, and robustness criteria throughout their lifecycle. See also Chaps. 8 and 10, and Sect. 18.3. And observe the definition of “substantial modification” in the AI Act. The definition is repeated in Annex B of this book.

Requirements

The rapid integration of AI into safety-critical systems necessitates robust standards to ensure functional safety. The European Union’s AI Act underscores the importance of developing such standards by setting high-level safety requirements for AI systems. However, the development of detailed, mature AI standards remains a challenge, as they are still incomplete and immature compared to standards in other domains; see also Sect. 1.2. Below, we present relevant AI standards (see also Chap. 8).

Generally, the AI Act emphasizes transparency, while standards focus more on explainability. ISO/IEC TR 5469:2024 provides guidelines on ensuring the functional safety of AI systems. It addresses the unique risks associated with AI, such as unpredictability and lack of transparency, by recommending comprehensive risk assessment and mitigation strategies. The standard emphasizes the need for continuous monitoring and validation of AI systems throughout their lifecycle to maintain safety integrity. The TR also includes an annex that analyses the applicability of techniques and measures presented in IEC 61508-3:2010 Annexes A and B to AI technology elements.

None of the standards and documents listed below are referenced in ISO/IEC TR 5469:2024. ISO TR 4804:2020. “Road vehicles: Safety and cybersecurity for automated driving systems. Design, verification, and validation”. This document explains how to create and test self-driving car systems using safety guidelines. It covers how to build these systems with safety and cybersecurity in mind, and how to check and confirm that the systems work correctly. The focus is on cars with advanced self-driving features (levels 3 and 4; see Chap. 17). It also discusses how to keep these systems secure while also making sure they are safe. This standard will be replaced by ISO/CD TS 5083.

ISO 24089:2023 “Road Vehicles—Software Update Engineering” standard focuses on software update engineering for road vehicles but can be adapted to other systems and domains, offering guidelines to ensure secure and effective software updates. It outlines how organizations should plan, develop, and implement updates for vehicles and electronic control units (ECUs), emphasizing the importance of maintaining vehicle safety and cybersecurity throughout the process. The standard provides steps for verifying, validating, and approving software updates before they are deployed. It also covers how to manage software update campaigns, including communication with users, handling update dependencies, and ensuring the integrity of the software.

ISO/IEC 23894:2023 offers guidance on risk management for AI technologies. It outlines a structured approach to identifying, assessing, and mitigating risks associated with AI applications. This standard is relevant for organizations seeking to integrate AI into their operations while maintaining a high level of safety and compliance with international standards.

ETSI GR SAI-007:2023 The document is one of the very few that identifies its target audience as designers and implementers who are making assurances to a layperson. The document identifies steps to be taken by AI platform designers and implementers to assure the explicability and transparency of AI processing, which includes AI decision-making and AI data processing.

IEEE P7001:2022 addresses transparency in all autonomous systems, focusing on those capable of causing harm. It includes both physical systems (e.g. automated vehicles) and non-physical systems (e.g. chatbots) and covers machine learning systems and their training datasets. The standard provides a framework for designing and reviewing transparency features, setting requirements, and methods for demonstrating conformance. Future standards may develop from this guideline to focus on specific domains.

NIST AI 100-1:2023 Framework (STD44), developed by the National Institute of Standards and Technology (NIST), provides a comprehensive approach to AI risk management. It emphasizes the need for a thorough understanding of the AI system's behaviour, rigorous testing, and validation to ensure safety. The framework also addresses ethical considerations, such as privacy and non-discrimination, which are increasingly relevant in AI applications. The document includes an Annex B describing how AI risks differ from traditional software risks.

Agile Adaptation

The development of AI systems fits very well with an agile and DevOps approach. If you already have an agile and DevOps approach, it is far easier to establish AI-important processes, such as training, frequent releases, and scaling. See also SafeScrum and DevOps in Chap. 10.

Safety Plan Issues

This depends strongly on the project, context, and the domain in question. We are entering new territory here, especially with high risk AI systems. It is crucial to have clear and detailed steps to ensure these systems are implemented safely and effectively. This includes defining specific procedures, safety protocols, and continuous monitoring to manage any potential risks. Therefore, ensuring that one has familiarity with the relevant standards and guidelines is a good start together with establishing a team of both safety and AI experts.

6.2 Explainable AI (XAI)

Objective

This chapter introduces how explainability of AI systems should be considered when developing AI technologies for the safety-critical domain, keeping in mind that this topic is under rapid development.

Information

This chapter has a direct link to the safety case part AI Safety Report (ASR) encompassing a subpart with the same title.

Explainable AI (XAI) is a research field receiving an increased interest in the last decade (Arrieta et al., 2020). However, both research literature and standards terms such as transparency, interpretability, and explainability are seemingly used interchangeably, challenging a common understanding of XAI (Vatn & Mikalef, 2024).

Explainability of AI Systems from a Technical Perspective

From a technical perspective, XAI might be considered a field concerned with creating "a suite of techniques that produce more explainable models whilst maintaining high performance levels" (Adadi & Berrada, 2018, p. 52138). These techniques are in many instances understood as something that is added on top of AI models that appear as black boxes to extract information about the inner workings of the model. However, explainable models could also be interpretable by design, meaning that the models are explainable without the need of external techniques (Arrieta et al., 2020). The models that need external XAI techniques to be explained could either use model-specific or model-agnostic techniques. The model-agnostic principle could be used for any AI model with the intention to extract some information about the inner workings of the AI model, while the model-specific techniques are only to be used for specific AI models (e.g. deep learning models). The different types of XAI techniques provide different types of explanations, separating between global and local explanations. While the global explanations seek to give an understanding of the overall prediction process of a model, the local explanations focus on explanations in response to a specific input. When developing AI systems, XAI techniques serves as important tools in the development process.

Explainability of AI Systems Considering Human and Organizational Aspects

Recent definitions place the audience as a key aspect when defining XAI (Arrieta et al., 2020; Adadi & Berrada, 2018; Vatn & Mikalef, 2024), underscoring the importance of considering XAI as a multidisciplinary field, which in addition to representing a suite of techniques also considers the recipients of the explanations of the AI systems. By adding both a human and organizational perspective to XAI, one is better equipped to understand how AI systems should be developed and explained to ensure alignment with knowledge on how the human brain processes information. During the design of high risk system, carefully considering this aspect is important to ensure that human operators can serve the controllability function and to avoid confusion and "out of the loop" problems (see Chap. 16). Also, considering organizational aspects is important when developing and implementing XAI techniques in high risk systems. Different stakeholder groups might have different sets of tasks, responsibilities, information needs, and decision areas attached to their role, and a careful analysis of this when implementing XAI will be important to ensure that people receive the necessary explanations to fulfil their role.

Requirements

The AI Act does not directly address XAI itself (Panigutti et al., 2023). However, the term transparency is given attention (article 13), as well as the principle human oversight (Article 14). Transparency of AI could be described as the opposite of opacity of AI, and in the context of the AI Act it is framed as a tool providing insight into the opacity of AI systems appearing as black boxes (Panigutti et al., 2023). High risk systems shall according to the AI Act—Article 13 be designed to be transparent so deployers can both interpret a system's output and use it appropriately. Article 13 also underscores that high risk AI systems shall have instructions for their use and lists what the instructions should contain as a minimum. There are several requirements, but it is possible to extract several dimensions relating both to the technical aspect of transparency and to the human aspect. For instance, the instruction for use shall include information about technical capabilities and characteristics of the high risk AI system relevant to explain its output. But also, information about the system's performance regarding specific persons or groups of persons on which the system is intended to be used shall be provided. This illustrates that in the context of high risk systems, transparency relates to explainability, as well as the human aspect.

ISO 5469:2024 "AI and functional safety" points at the degree of explainability as an important consideration when assessing risk factors of AI systems and distinguishes between transparency and explainability. While explainability is defined as "the property of an AI system to express important factors influencing the results of the AI system in a way that is understandable to humans" (p. 15), transparency is defined as "the property of a system that appropriate information about the internal processes of an AI system is made available to relevant stakeholders" (p. 15). For system developers, important considerations will be to find the sufficient degree of transparency and explainability. Important questions for developers of high risk AI systems will relate to whether sufficient information about the system is available, how this information should be conveyed to a given recipient, and whether it produces complete and correct results that are reproducible. ISO 5469:2024 also points at explainability approaches that can be used to provide interpretability or explainability into the model structure that might prove helpful in verification and audit processes.

ISO IEC 23984:2023 "AI and risk" underscores how inclusion of stakeholders in development of AI system can help to identify the goals and describe means for enhancing transparency and explainability of AI systems. Lack of both transparency and explainability is listed as a source of risk; however, excessive transparency and explainability are also mentioned to be a potential source of risk relating to confidentiality, intellectual property, and security.

ISO IEC TS 8200:2024 "Controllability" is a relevant standard to consider in relation to explainability of AI systems. Controllability is important to accommodate the principle of human oversight in the AI Act. While degree of explainability might be considered important for a system's controllability, ISO IEC TS 8200: 2024 also points at how controllability might enhance the trust in an AI system that is not fully explainable.

DNV has published a recommended practice DNV RP 0671 "Assurance of AI-enabled systems", which underscores the requirement for evaluating the explainability during the assurance of the trustworthiness of an AI system. The RP also distinguishes between interpretable models that can be understood through its design and models that need additional techniques to be explainable.

Agile Adaptation

The development of AI systems accommodating the requirements for transparency and explainability fits very well with an agile DevOps approach. Considering the evident user aspect in those cases a human is supposed to serve as the controllability function of an AI system, user-centred design principles used in combination with a DevOps approach would be very useful (see Chap. 16).

Safety Plan Issues

When developing high risk AI systems, it is important to have the right balance between transparency, explainability, and maintaining system confidentiality. Different stakeholders, such as manufacturers and users, will require varying levels of detail, so a clear strategy for how and when to communicate these aspects is important. Careful consideration must also be given to ensure that explainability supports the human oversight function without causing confusion or leading to "out of the loop" issues. Regularly updating XAI techniques and refining their alignment with stakeholder needs should be part of the agile DevOps process.

6.3 Safety Of The Intended Functionality (SOTIF)

Objective

The main objective of the SOTIF plan is to describe the planning activities and their justifications to ensure that the risk level associated with hazards related to intended use is taken care of, i.e. are sufficiently low. Related to this, it is important to consider three categories of data—testing data, verification data, and training data.

Acceptance criteria and validation targets are critical for SOTIF safety. Acceptance criteria specify acceptable levels of harm, and validation targets measure efforts to meet these criteria. These should consider relevant operational design domains. Note that UL 4600 lists SOTIF as one of 16 highly recommended hazard identification techniques.

Information

Generally, verification testing helps to ensure the software meets the specified requirements and standards. In contrast, validation testing ensures that the software meets the needs and expectations of the end users (Tank, 2023). See also Chap. 9

about testing in this book. In addition, when dealing with AI systems, we have to consider the training data which are used to “teach” the AI system how to respond to each input and system state. Acceptance criteria, based on verification testing and validation targets, are the most important information used to assure the safety of the intended function (SOTIF) of a system. Acceptance criteria are often defined as an acceptable number of fatalities, injuries, or property damage events in a certain number of hours of operation. Validation target, on the other hand, is the amount of effort required in terms of hours of operation to show that the acceptance criteria are met. See also Chap. 8: “Planning the safety activities: tests, analysis, scenarios, verifications, validations, and regression”. It is our experience that descriptions of acceptance criteria and validation targets often overlook factors such as operational design domain (ODD) and operational lifetime.

When it comes to hazards, ISO 21448 defines four sets of scenarios (see also Chap. 9), which are elaborated in the standard:

- Known, not hazardous scenarios
- Known, hazardous scenarios
- Unknown hazardous scenarios
- Unknown, not hazardous scenarios

The safety case arguments must include arguments that all of the four categories of hazardous scenarios mentioned above are handled in a safe way. The SOTIF standard ISO 21448:2022 describes how it could be done.

The safety case needs to handle information related to training data, validations data, and testing data. An important part of the safety case is arguments that the validation data and testing data mirror the needs and requirements of the system’s users. For an AI system we also need arguments that the training date is appropriate.

Requirements

As mentioned earlier, we need links between the software product and the test data, validation data, AI training data, and the ODD description. Changes to one of the datasets must be reflected in the other datasets. This will require a large set of links of type “this part of the validation data is dependent on this part of the ODD”. In addition to maintenance of the datasets, we will also have the need to maintain all relevant document links.

Agile Adaptation

Changes can and will occur in any software development project. What is special for an agile project is that we claim to embrace changes, also in the ODD. For a SOTIF analysis, this poses an extra challenge since changes to the system’s functionality will require extra safety analysis. In addition, it might invalidate some of the current ones.

As a consequence of this, we need a project culture and way of working that always is able to handle changes to the environment and requirements in an efficient manner. If we do this, we are able to start early in the project with a rather imprecise definition of the ODD and the intended functionality. We will then be able to adapt as our ODD understanding increases without introducing errors due to outdated information or misunderstandings.

Safety Plan Issues

Planning for change implies that we document all relevant links/references between all relevant documents—especially links between:

- ODD, test data, validation data, and training data
- Requirements, test data, and validation data
- ODD, requirements, and software product

As we learn more about the ODD—users, the operating environment, and so on—the amount of unknown scenarios, hazardous or not—will shrink. This will improve the possibilities for relevant safety analysis and thus increase our confidence in the analysis of the safety of the intended functionality of the system.