DGUV Test
Prüf- und Zertifizierungssystem der
Deutschen Gesetzlichen Unfallversicherung

05

DGUV Test Information

Date of Issue: 04/2021

General Principles for Assessing the Safety of Artificial Intelligence (AI)

This document specifies general principles for assessing the safety of AI technologies. However, it does not address ethical or liability issues. The principles define the requirements for AI technologies with regard to health and safety and serve as a guide for the development of product-specific test requirements.

The term AI has not had any standard definition so far but has served primarily as a generic term that covers a variety of procedures such as logical reasoning, expert systems, various search algorithms and machine learning. With the development of the new ISO IEC 22989 standard, a uniform definition for AI and machine learning is to be established in the future. In this DGUV Test Information, AI and machine learning are defined as follows:

Artificial intelligence

The ability to acquire, process, create and apply information about objects, events, concepts and rules as well as their relationships and properties, where they are organised for the purpose of systematic, target-oriented use and where they are available in the form of a physical, mathematical or otherwise logical presentation of a system, entity, phenomenon, process or data, with a view to carrying out one or more particular activities in order to achieve a specific goal.

Machine learning

Process using computer-assisted technologies in order to enable systems to learn from data or experience.

Principle No. 1
If a task can be carried out with the help of conventional technology, then that technology should be given preference over the use of AI.

This position can only reflect the current state of the art and must be adapted or, if necessary, abandoned as technology advances. Today, the requirements for functional safety in industrial applications are specified in the safety standards ISO 13849-I, IEC 62061 and IEC 61508.

Due to the high performance requirements on plants and machinery in this context, the hazard potential that is involved in handling them is high or even very high.

To control such hazards, safety functions which are designed as two-channel redundant systems are often used today. In addition, the applicable safety standards place specific requirements on the quality of the relevant components concerning reliability, fault detection (diagnostic coverage) and the reliability of the software.

The aim of the measures is to achieve so-called single-fault safety. If a fault occurs in one channel of the safety function, then safety still continues to be available via the second channel, and fault detection then leads to a safe state. Neither ISO 13849-I, IEC 62061 nor IEC 61508 currently contain specific requirements for the application of AI.

Today, the main components of AI are algorithms, realised, for example, via neural networks that have been trained for their tasks or objectives with the help of suitably prepared data sets. Most AI applications are currently based on machine learning procedures. The aim of machine learning is to create a model based on training data via a training process, and this model should then be able to generalise knowledge and become applicable to new data. During the training process, the model is optimised by an algorithm. In the case of *deep learning* (e.g. deep neural networks), this model is usually highly complex and therefore difficult to understand. This makes a safety assessment considerably harder, as it requires the use of procedures for the verification and validation of black box systems.

Principle No. 2
Where AI is used for the realisation of assistance systems, it provides support for humans, but it cannot be rated as a safety function.

Today, technical assistance systems are used in many areas of industry as well as in everyday life. Research Report 502 of the German Federal Ministry of Labour and Social Affairs (BMAS) distinguishes between physical, sensory and cognitive human-related assistance approaches.

Whereas mainly physical assistance systems have been used until now, the new intelligent AI-based assistance systems can support the cognitive and sensory capabilities of an operator. The assistance system, which is equipped with sensors matching its purpose, monitors the ongoing process at the same time as the operator and then responds independently to events and to changes in process parameters. Here, it is important to understand that, where personal protection is concerned, assistance systems do not usually meet the requirements of the relevant safety standards – such as ISO 13849-I – and must not therefore be considered as safety functions. Basically, the operator shall always have the option of gaining full control over the process.

Physical assistance systems

provide support for demanding physical activities. They also compensate for diminishing physical capabilities and protect against their premature loss. The current state of the art ranges from mechanical/motorised power assistance and personalised assembly workstations for simple, rule-based work situations to adaptive, collaborative robotic systems for complex, highly variable, and expertise-based processes in production, assembly, and maintenance. This primarily involves supporting the musculoskeletal system and sensory organs.

Sensory assistance systems

compensate for functional, often age-related, changes in sensory organs. Advanced systems primarily address hearing- and vision-related limitations and provide combined cognitive-sensory support (e.g., augmented reality glasses).

Cognition-supporting assistance systems

primarily provide application-oriented, near-real-time information, thus supporting the workforce in making decisions. Depending on the level of support, such functional support mainly concerns the ability to respond, think, remember, and draw conclusions (cf. Müller et al. 2014). The main hardware elements in application-oriented research projects are mobile devices and interactive visualisation systems.

Source: Forschungsbericht 502 BMAS (Research Report)

Principle No. 3
Continuous learning systems shall not have any dangerous impact on safety functions.

Depending on the design, a machine learning system can either stop its learning process, so that it always behaves in the same way on all subsequent occasions, or it can continue to learn while it is being used. This is the approach followed by continuous learning systems. They use the incremental ability of the AI system, which is continually active throughout the operational phase of the system lifecycle.

Whereas the behaviour of non-continuous learning systems is fixed throughout the development process and is not meant to change during the operational phase, continuous learning involves a step-by-step updating of the model during the operational phase. In such a case, any data that is acquired during the operational phase is not only analysed with a view to producing an output, but it is simultaneously also used to adjust the model within the system, so that it can be improved on the basis of the additional input data.

The aim of continuous learning is to rectify problems and errors caused by training data that were initially severely limited or incomplete – or to respond to gradually changing operating conditions where they differ from the specified training environment, thus mitigating the problem of nonconformance to the specified concept.

By applying incremental adjustments to the model, continuous learning causes the AI system to behave dynamically. On the one hand, this is intentional, but, on the other, it also poses considerable challenges. It is very difficult to ensure the correct functioning of such a system during the operational phase, as this also requires continuous system verification. Furthermore, the new input data would have to be recorded, so that any later factory update can integrate the data into the new training record or so that the new input data can be used for the purpose of troubleshooting. It follows that each application must involve weighing up the pros and cons of this method and any risks that might arise.

Principle No. 4
The “decision-maker” shall be safe.

A decision means choosing an action among two or more existing potential alternatives with due consideration of certain higher-level goals. Each activity requires decisions. If, for example, the safety-oriented behaviour of a machine depends on this decision, then it shall be made quickly, yet with careful consideration. In general, a distinction is made between a human and a technical decision-maker. In automation, safety decisions are usually made by a control, i.e., by a technical decision-maker. This should also be given preference based on the hierarchy of measures (technical measures preferred over organisational measures). The role of the technical decision-maker is to keep the AI-controlled system in a state where the risk to the operator remains acceptable.

This can be, for instance,

- a degraded state (equivalent temporal state, applying alternative measures), or
- a previously specified and defined state.
- Alternatively, this can lead to a controlled or
- uncontrolled shutdown.

The control mechanism can only make such decisions with the help of AI if certain boundary conditions are met, e.g., requirements on data quality, data storage, hardware and software.

Principle No. 5
Data quality shall be monitored and ensured.

Reliable data and its processing are essential requirements for correct decisions. To ensure high data quality, suitable components, e.g., redundant sensors, shall be used for the collection of data, in compliance with the relevant product standards. In addition, any data that has a significant impact on the decision shall be monitored at certain intervals and checked for plausibility. Incorrect or incomplete information can negatively influence the decision.

Note: Data quality also includes the quality of the entire data set, i.e., in terms of completeness and diversity. This is to ensure that all the relevant modelling examples are available in the data set.

- **Example 1:** A data set intended for the creation of a personal identification system should contain sample images of persons differing in height, shape, skin colour and facial features, as the system cannot otherwise recognise diversity among individuals.
- **Example 2:** Periodic fault detection regarding safe and reduced speeds of machine tools (monitoring of the encoder)

If the values of the encoder are also given plausibility checks during the process, then faults are detected at the time of occurrence. The data quality is thus permanently ensured.

To allow the further processing of data by a decision-maker, the data needs to be output in a readable format, and its integrity shall be ensured. This can be done, for instance, via modern bus systems.

Principle No. 6
Relevant decision-maker's data shall be recorded and stored.

When making a decision, all data used by the decision-maker should be given a timestamp and saved for a certain period of time, so that the decision can be traced. The parameters that are relevant for this purpose must be specified before the AI system is put into operation.

Principle No. 7
The decision-maker shall meet the design principles with regard to functional safety.

Safety-related decisions need a suitable hardware and software structure, e.g., to allow a detailed diagnosis (detection of faults and the current state). Therefore, the decision-maker shall be designed on the basis of a functional safety standard suitable for the area of application. Each decision shall comply with the safety level identified in a risk assessment.

The diagnosis and decision-making shall, in particular, take place against the background that faulty states can be caused by system failures or common cause failures, not just by random component failures. Such failures can negatively influence the decision.

A technical decision-maker requires a high level of safety in the decision-making process. It is especially important that the decision-maker detects whether the issue is a random component failure or a system fault. The design principles also include the application of specified activities concerning functional safety management.

Note: Concerning machine safety, for example, ISO 13849-I is to be applied.

Principle No. 8
AI technologies shall not suggest a level of safety that does not exist.

As shown in Principle No. 1, AI-based systems are not a substitute for classical protective devices for personal protection as defined in ISO 12100 according to the current state of the art.

Where AI-based systems intervene in machine or vehicle controls, whether directly or in an assistant manner, they shall not suggest to operators a level of reliability or safety that is not guaranteed through technical measures (Principle No. 2). The manufacturer shall ensure that the safety limits of the system are clearly communicated to the operator. Wherever this is possible, such limits shall be clearly indicated to the operator during use, e.g., whenever he approaches those system limits in his behaviour.

The following principles shall be observed:

- The manufacturer shall provide unambiguous descriptions of the functions and the performance limits of the system.
- Safety level of hard- and software: The higher the level of automation and risk, the higher the safety level required.
- The functional status shall be displayed to the operator.
- Self-monitoring / fault diagnosis: Depending on the criticality of a fault, a safe state shall be initiated.
- The design of the human-machine interface (HMI) shall meet ergonomic requirements.
- It shall be possible for humans to override or shut down the system.

Principle No. 9
Security (protection of the AI system against manipulation) shall be ensured.

When using the AI systems, sufficient (personal) protection shall also be ensured at each stage of the product lifecycle.

Neither the AI within the system nor unauthorised external access may modify the functional safety of the overall system in such a way that personal protection is no longer ensured. For this purpose, all safety-related parts of the control system (functional safety) shall take priority over process control, including AI, and shall be protected against manipulation (whether through AI or external access). This includes the appropriate fulfilment of IT security requirements as specified in the IEC 62443 standards and covers the requirements on individual safety components (IEC 62443-4-2, DGUV Test Principles GS-IFA-M24) and on the overall system.

Principle No. 10
The field behaviour shall be observed.

AI has the advantage that it generalises common cases and that it can therefore be used for the solution of complex tasks in complex environments. However, this raises the issue that neither the anticipated use nor the anticipated environment can be fully specified in advance.

If, for instance, a self-driving vehicle is intended for use in road traffic, its anticipated working environment cannot be specified completely, as this would require a detailed and continually up-to-date description of all roads, including their environments. Similar difficulties arise with systems for detecting persons, as persons have a high level of visual diversity.

During the specification stage, an analysis is therefore required, examining the intended application and the intended working environment and elaborating the instances that shall be covered by the future model in all cases. In particular, this must include the identification of rare instances. However, the high level of complexity means that even a far-reaching analysis, conducted in advance, can miss some exceptional situations. Also, the anticipated working environment can change over time.

This makes it necessary to review the impact of the algorithmic system at regular intervals. A manufacturer shall therefore set up, maintain and regularly update a system that collects and reviews information about the product. This system shall contain information from the implementation and post-implementation phases as well as all publicly available information on similar products in the market.

Having collected this information, it should then be assessed for its potential relevance to the AI trustworthiness of the product. In particular, it is important to assess whether there are any previously undetected risks and whether any of the risks assessed so far have ceased to be acceptable.

Published by:
Deutsche Gesetzliche Unfallversicherung e. V. (DGUV)
German Social Accident Insurance (DGUV)
Glinkastraße 40 · D-10117 Berlin
Telephone: +49 30 13001-0 (switchboard)
E-mail: info@dguv.de · Internet: www.dguv.de

Contact:
Secretariat DGUV Test
Alte Heerstraße 111 · 53757 D-Sankt Augustin
Telephone: +49 30 13001-4566
E-mail: dguv-test@dguv.de

Available:
www.dguv.de/publikationen Webcode: p022249

Further information:

We test for your safety.