Safe machines with – or despite – artificial intelligence[1]

Mattiuzzo, Corrado (Commission for Occupational Health and Safety and Standardization (KAN), Sankt Augustin); Vock, Silvia; Mössner, Thomas; Voß, Stefan (German Federal Institute for Occupational Safety and Health (BAuA), Unit Workplaces, Safety of Machinery, Operational Safety, Dresden)

In April, the European Commission produced a proposal not only for a regulation governing artificial intelligence, but also for a regulation concerning machinery products. This latter regulation, which is to include binding framework conditions for the use of artificial intelligence, is to replace the Machinery Directive 2006/42/EC. The task is now to review whether these framework conditions contain complete, clear and verifiable requirements setting out in what cases and subject to what criteria safety-related functions of a machine may be performed automatically by artificial intelligence methods or under the influence of such methods. This article aims to provide corresponding information and suggestions.

I Statutory framework

In the European Union, manufacturers of a machine are required to assess the risks presented by it and reduce them as far as possible. For this purpose, they must:

- clearly specify the intended use of the machine, and anticipate possible forms of misuse which might reasonably be foreseen;
- eliminate the associated hazards or mitigate the risks associated with them according to a defined priority;
- take account of the severity of possible injuries or harm to health, and the probability of their occurrence.

A further requirement is that during the machine's entire anticipated life, it must not present a risk greater than that determined as being acceptable by the risk assessment performed before the machine is placed on the market. It follows that in combination, the two new items of legislation must ensure that the provisions set out within them concerning risk assessment/risk management requirements adequately reflect this objective.

It is therefore crucially important that manufacturers are able to assess the risks presented by their products. This is precisely the challenge that would emerge if, for example, a control system supported by machine learning were to be relied upon to prevent people from being endangered by moving parts of a machine: the designers of systems based on the more complex artificial intelligence methods

[1] *The original article in German was published in issue 6/2021, page 188, of the journal "Arbeitsschutz in Recht und Praxis"(ARP, https://rsw.beck.de/zeitschriften/arp), published by C.H.BECK. Translation by the KAN Secretariat.*

(such as machine learning employing neural networks) have as yet often been unable to explain satisfactorily, even after the event, why their systems behaved in a certain way. Moreover, it is difficult to demonstrate that the learnt model is correct, not least because the data used for training represent only a subset of all possible input values. The risk of certain inputs during the system's life leading to incorrect decisions cannot therefore be reliably excluded.

In some cases this challenge may arise even on highly complex transparent and comprehensible models, such as decision trees, with the consequence that their results cannot be evaluated in advance by the use of traditional methods.

II The role of machine control systems

Where control systems are used to execute safety functions, they have a significant bearing upon the safety of a machine. Modern, automated machines frequently execute applications in the absence of any direct human action, and are connected to their environment and/or other devices by sensors and actuators. To enable these machines to take the necessary decisions, action strategies are programmed into their control systems. These strategies may be simple or highly complex, irrespective of whether they are based on traditional software or artificial intelligence methods.

Depending on their complexity, the programmed action strategies and functionalities may lend themselves to verification and assessment by means of existing, proven procedures. This also holds true for artificial intelligence methods of lower complexity. The technical principles and assumptions upon which traditional good-practice methods for assessing system safety engineering are based are however not suitable for more complex artificial intelligence methods. One example of this is that until now, the occurrence of random faults was assumed only for hardware components of control systems, and software component failure was always ascribed to systematic faults. This concept is no longer tenable for the majority of machine learning (ML) methods. For example, an ML algorithm may continue to learn from the data acquired during operation and thereby adapt to new conditions. As a result, the underlying programs and action strategies of the control system are no longer defined in static and comprehensive terms; faults may then be caused not only by incorrect programming, but also by an incorrect strategy, i.e. one which has been learnt by the control system but is not consistent with the goal of development. Equally, even static and thus defined action strategies in the control system, such as those arising in ML methods that do not continue to learn in running operation, may also present challenges for assessability and verifiability.

The implications of the action strategies and functionalities discussed above can thus be classified in relation to machine safety as follows:

Case 0: No implications for machine safety

The design of the control system excludes, with the reliability[2] required by the risk assessment, the possibility of the decisions based on the action strategies having direct or indirect impacts upon safety.

Case 1: Can be solved without limitations under the Machinery Directive[3]

The decisions taken by the control system which have implications for safety depend solely upon action strategies which can be assessed by means of methods that have already been proven for such action strategies.

Case 2: Can be solved only to a limited extent under the Machinery Directive

The decisions with implications for safety taken by the control system also depend upon action strategies which cannot (yet) be assessed by means of methods that have already been proven.

Case 2a: Consistent with the requirements of the Machinery Directive under certain conditions

The design of the machine rules out, with the reliability required by the risk assessment, the possibility of decisions based upon the action strategies having impacts which increase the risk beyond that deemed by the risk assessment to be acceptable, or giving rise to new risks.

Case 2b: Not consistent with the requirements of the Machinery Directive

The design of the machine is not able to rule out, with the reliability required by the risk assessment, the possibility of safety-related decisions based upon the action strategies having impacts which increase the risk beyond that deemed by the risk assessment to be acceptable, or of giving rise to new risks.

[2] Probability of satisfying the condition under given conditions for a given time interval

[3] Where reference is made in the text to the Machinery Directive, it refers to the current Directive 2006/42/EC and the authors' aspiration for the machinery products regulation proposed by the Commission on 21 April 2021 , and finally to be agreed by the European Parliament and Council.

III Example cases

Examples for case 0 from ISO/TR 22100-5[4]:

Optimized packaging (ISO/TR 22100-5, 4.2.1.1)

A robot which is optimized by machine learning loads a pallet with parts varying randomly in their size. In this case, the design of the machine assures that the predefined dimension or weight limits are observed, thereby reliably excluding the possibility of the loading strategy optimized by machine learning giving rise to additional hazards or exacerbating existing hazards. For this purpose however, the risk assessment must for example include demand rates for the downstream safety functions (such as for observance of the dimension and weight limits) which cannot be exceeded by the loading strategy.

Optimized spraying of herbicides (ISO/TR 22100-5, 4.2.1.2)

Image processing with the use of AI machine learning methods enables crops and weeds to be identified more precisely, permitting more precise decisions on where and in what quantities herbicide is to be sprayed. For this purpose, the machinery used for application of the herbicides is equipped with cameras connected to the spraying nozzles. Based upon the detected image, the ideal quantity of herbicide is applied through the individual spray nozzles in the locations at which weeds are detected. A cab (with a filter or overpressure system) on the spraying machine itself or on the agricultural tractor towing it ensures that the sprayed herbicides do not present a health risk to workers. It is ensured that the use of machine learning to optimize identification of the weeds does not give rise to hazards additional to or greater in scale than those of the conventional machine function.

Optimized retrieval of parts from a laser cutting machine (ISO/TR 22100-5, 4.2.1.3)

A machine uses laser beams to cut parts fully automatically in an almost infinite range of geometries, sizes and thicknesses from metal sheet, and retrieves them with the aid of 2,500 suction cups and 180 pins. Should retrieval not be successful at the first attempt, the machine uses alternating methods optimized by machine learning until it is successful in pressing the part out of the scrap skeleton. Guards prevent access to the cutting table and part retrieval point. The retrieval strategies optimized by machine learning neither cause additional hazards nor exacerbate existing hazards in comparison with the conventional machine function. Here too however, demand rates for these guards (e.g. access by personnel in order to clear blockages caused by parts that have not been retrieved) must have been considered by the risk assessment. The retrieval strategies must not violate these demand rates.

[4] ISO/TR 22100-5:2021-01, Safety of machinery – Relationship with ISO 12100 – Part 5: Implications of artificial intelligence machine learning

Examples for case 1:

Complex traditional software in a safety component, assessable in accordance with the concepts of functional safety

A danger zone is monitored by means of optical sensors. Software is used to combine the image data acquired by three sensors spaced apart from each other into a three-dimensional image of the danger zone and to monitor the zone for the intrusion of objects. Should an object enter the danger zone, a safe stop is triggered. Proven methods were used to develop and assess the complex safety-related software.

Less complex artificial intelligence method that can be assessed by proven methods

The safety-related parameters of an aluminium die-casting plant are registered by sensors and evaluated in real time in order to shut the plant down should the pressure or temperature lie outside the permissible safety range. A higher-level anomaly detector employing a decision-tree algorithm detects faults in the sensors, network or hardware and triggers a safe stop. The supplementary use of the anomaly detector enables redundancies in the safety system to be reduced with no impairment to the safety integrity of the system as a whole. Since anomaly detection employs a model that lends itself to explanation and interpreting and the algorithm is used as a fully trained model over the system's life with no further learning during running operation, the software of the anomaly detector can be developed, verified and validated by means of recognized software quality management methods. This also enables proven methods to be used for risk assessment of the die-casting plant.

Examples for case 2a:

Driverless transport vehicle system consistent with the Machinery Directive (inspired by but deviating from ISO/TR 22100-5, 4.2.2)

A driverless transport vehicle system operates within an area without access safeguards but with clear delimitation, and optimizes its navigation autonomously by machine learning. Collisions are avoided by technical measures employing protective sensor devices and speed adjustment programmed conventionally. The requirements for collision avoidance are much greater in this case than those for driverless transport vehicle systems operating only in designated lanes, and greater still when compared to systems with access safeguards. However, no new hazards are created, and the risks can be assessed and reduced to an acceptable level by means of proven methods.

Supplementary signalling assistance systems

A large machine tool features state-of-the-art protective equipment in the form of guards and sensors in order to protect operators and third parties. In addition to these measures, which under current product safety legislation are already adequate, an assistance system provides an acoustic warning when a person enters the danger zone without itself intervening in the machine control system. The assistance system continues to learn during running operation of the machine and is thereby able to optimize detection of persons continually. Practical experience useful for possible future applications is also built up in this way[5].

Example for case 2b:

Automated driverless transport vehicle system not compatible (at present) with the Machinery Directive (inspired by but deviating from ISO/TR 22100-5, 4.2.2)

A driverless transport vehicle system operates within an area without access safeguards but with clear delimitation, and optimizes its navigation autonomously by machine learning. Collisions are avoided by technical measures employing protective sensor devices and, in deviation from example 1 in case 2a, by speed adjustment optimized by machine learning. The machine learning method used lends itself neither to interpretation nor to explanation, and continues to learn dynamically during running operation of the system. Here too, new hazards do not necessarily arise; it is however no longer possible for the risks to be assessed by proven methods and thus safely reduced to an acceptable level. Owing to this lack of means for assessment of the risks, the case constellation described is not consistent with the requirements of the Machinery Directive.

IV Conclusions

On the one hand, approaches exist by which safety can reliably be demonstrated even of highly complex technical constellations which cannot currently be assessed. This is possible by the definition of "arguments" which are intended to provide strong circumstantial evidence obtained by inductive reasoning (but not absolute proof). Such approaches have long been used for example in nuclear technology or aeronautics and aerospace, and also for determining whether software is suitable for safety-related use.

Attempts are now being made to use such approaches, which tend to have their origins in the field of risk management, to create catalogues of criteria for attainment of an acceptable level of risk that can also be applied to methods of artificial intelligence. These criteria may concern specification and modelling, explainability and accountability of decisions, transferability to different situations, verification and validation of the system, monitoring during the runtime, human-machine interaction, process assurance and certification, and also safety-related ethics and data security. This is evidently the direction taken by the European Commission's proposal on artificial intelligence. Under an approach of this kind,

[5] The example does not address aspects of possible effects on operator behaviour, which are not considered in the context of this article.

safety is defined primarily not by verifiable product properties, but by verifiable process criteria.

At the same time however, in order to attain a level of safety approximating that embodied in the European product safety regulations and the basic principle of prevention at the workplace, the criteria for the aforementioned "arguments" must first be shown to be complete and reliable. If a conservative approach is adopted, regulations which are intended to set out the framework and basic requirements for this purpose cannot therefore be formulated until the assumptions on which they are based have been reliably proven. In the view of the European legislators however, the time required for this purpose was not available: the Commission proposals for revision of the Machinery Directive and for artificial intelligence have now been made available simultaneously.

Whether the two proposed regulations contain complete, clear and verifiable requirements setting out in what cases and subject to what criteria safety-related control functions of a machine may be influenced or determined automatically by artificial intelligence methods, or whether an adequate legal footing for procedures and assessment strategies for proper risk assessment are available, must now be reviewed. Should, at the end of the legislative process, the two items of legislation not satisfy these requirements, their application would lead to considerable uncertainty for the market players involved. It is clear that a need for interpretation, research and standardization still exists, and that the interdisciplinary approach to the subject poses major challenges for the work of the individual disciplines.

V Outlook

Both the essential health and safety requirements contained in the draft of the new machinery products regulation and the requirements contained in the draft of the artificial intelligence regulation are formulated as technology-neutral objectives of protection. Specific provisions for satisfying these objectives of protection will be set out in the harmonized standards supporting the two future regulations.

The following premises are relevant to application of action strategies for safety-related control functions:

- The safety-related control functions based on the action strategies must not lead to the limits defined by the risk assessment being violated.
- The safety-related control functions based on the action strategies must be executed with the reliability specified in the risk assessment for the safety function concerned.
- Methods must exist by which the reliability of the safety-related control functions based upon the action strategies can be assessed as part of the risk assessment.

Requirements for software for safety functions can be found in the EN 61508 series of standards[6] and in EN 62061 (last amended 2015-08)[7]. Requirements for software intended for use for safety functions in machines have now also been included in the draft of EN ISO 13849-1:2020[8]. It must therefore be determined whether these standards are sufficient and can be used for the development of software employing artificial intelligence.

Since safety functions make a not inconsiderable contribution to risk reduction, high demands are also made of their reliability, which is expressed by the required Performance Level (PL) or Safety Integrity Level (SIL). Where artificial intelligence methods are to be used in software for safety functions, they must of course attain the required high reliability values. A need still exists in this context for the reliability of the AI methods to be increased. At the same time, methods for checking this reliability must be developed the quality of which is sufficient to meet with general acceptance and is suitable for standardized checking of the reliability of safety functions with an AI component.

The following criteria would have to be met in order for example 2b to be considered consistent with the Machinery Directive:

- The software for the safety function of speed adjustment, which is optimized by means of machine learning, can be developed by the use of a method that is recognized, for example in harmonized standards.
- The reliability of the AI algorithm lends itself to checking, for example by means of harmonized standards.
- The safety function attains the reliability (PL, SIL) required of it even with the AI component included.
- A check following an optimization step ensures that the limits defined in the risk assessment are not violated.

It must also be clarified to what extent existing risk assessment methods are suitable for properly assessing artificial intelligence methods in conjunction with relevant application scenarios. In this context, a need exists for risk assessment methods to be developed further in order for the existing high level of protection of the European Single Market to be maintained, now and in the future.

[6] EN 61508 series of standards: Functional safety of electrical/electronic/programmable electronic safety-related systems

[7] EN 62061 (last amended 2015-08), Safety of machinery – Functional safety of safety-related electrical, electronic and programmable electronic control systems

[8] EN ISO 13849-1 (2020): Safety of machinery – Safety-related parts of control systems – Part 1: General principles for design