Legal framework for the deployment of autonomous and AI systems

Thomas Jürgensohn[1], Christina Platho[1], David Stegmaier[2], Matthias Hartwig[2], Mathilde Krampitz[2], Lorenz Funk[2], Timon Plass[2], Heiko Ehrlich[3]

baua: Focus

The use of industrial machines with AI-based algorithms seems to be possible. However, the impact of AI systems on the safety of employees and possible changes to the existing legal framework have not yet been systematically investigated. In the project F2432, a taxonomy was developed that identifies safety-relevant factors in AI systems in the industrial sector. Based on this taxonomy, recommendations for the further development of product safety law are given. By introducing different legal constructs, the legal framework can be adapted to the specific characteristics of AI systems.

Contents

1 Introduction

The goal of the project was to identify necessary changes to the existing legal system such as the product safety law and the industrial safety law necessitated by the introduction of AI-based algorithms and other software with autonomy features in physical systems. The overview is limited to industrial systems for which a safety assessment is required. These potentially hazardous physical systems, whose behavior is determined by software that includes AI-based algorithms, are referred to as "software-physical (AI-based) systems". Since AI-based algorithms have so far rarely been used in industrial areas such as automatization, robotics or mechanical engineering, highly automated (autonomous) motor vehicles were also adressed.

[1] HFC Human-Factors-Consult GmbH., [2] IKEM Institut für Klimaschutz, Energie und Mobilität e.V., [3] TÜV Nord

In a first step, it was specified whether these new systems might pose additional risks to employees and consumers that have not yet been taken into account from a legal point of view or that can no longer be handled by existing regulations. In a second step, it was examined which legal framework is required to guarantee safety when facing the identified potential hazards. It was also analysed whether the allocation of responsibility among the parties involved (manufacturers, operators/users, employees) according to preventive regulatory law (especially product safety law and industrial safety law) or repressive liability law can still be applied to these new systems.

2 Development of a taxonomy for categorising safety-related factors

In order to be able to answer these central questions of the research project, a system of categories (i.e. taxonomy) was developed, which summarizes safety-related factors in recently developed software-physical (AI-based) systems. The taxonomy was based on extensive expert interviews and supplemented by literature analysis. The expert interviews were carried out over a period of more than one year and were conducted as one-to-one interviews lasting 1-2 hours. Some of the 33 experts coming from the fields of robotics, smart home, AI, automatisation, automotive, and security participated in more than one interview. A notable percentage of the sample were experts in functional safety and active in various standardization committees dealing with AI-related safety aspects. However, only a few of the experts had dual expertise in both AI and security assessment. The field of AI and safety is, unlike the field of AI in information technology, still very young and has hardly been dealt with outside the area of highly automated vehicles. This is why only few experts had profound expertise in both AI and safety assessment. Up to now no field of knowledge concerning AI-related safety has been established.

The taxonomy was developed with industrial applications in mind; it might be possible to transfer it to other application areas - e.g., robotics in the smart home or medical care -, though area-specific modifications will most certainly be required.

3 The dimensions of the taxonomy

The taxonomy describes a total of 40 characteristics (factors or properties) of physical systems and their environments that influence or are related to safety. These characteristics are organized into seven dimensions and 14 subcategories. The top level of the taxonomy is formed by the seven dimensions "changeability", "transparency", "interconnectedness", "controllability", "resilience", "human involvement", and "harmful consequences". Each dimension summarizes two subcategories, which include between two and four characteristics, such as: "Ability for specification", "understandability", "predictability" or "autonomy". An overview of the taxonomy is given in Table 1.

Tab. 1 Graphical overview of the taxonomy

Changeability	
System	No changes after delivery
	Predictable and controllable changes after delivery
	Adaptivity (i.e. minor changes in few parameters during operational use)
	Major changes in operational use
Environment	Minor changes in controlled surroundings
	Unpredictable changes in complex surroundings

Transparency	
Experts	Ability for specification
	Definability of the limits of capability
	Understandability
	Predictability
Parties involved	Knowledge of system‘s scope and limits
	Predictability of system‘s state or actions
	Comprehension of system‘s functionality

Interconnectedness	
Internal	Centralized network
	Distributed network
External	Defined data (informative or action-guiding)
	Non-defined data

Controllability	
Emergence	Autonomy (relative independence from quality criteria or goals)
	Self-organization
Constraints	System constraints via virtual safety fencing or conventional systems
	Constraints in surroundings or application area
	Constraints in/control over data flow

Resilience	
Robustness	Stability (bounded-input-bounded output)
	Handling unknown situations or unforeseen events
	Security
Failure handling	Passive mitigation of consequences
	Active mitigation of consequences

Human involvement	
Human as agent	Humans as both a safety-ensuring and error-prone part of the process chain
	Misuse against intended use
	Deliberate disruption of the intended function or damage
Human put at risk	Instructed employees
	Users
	Consumers
	Third parties

Harmful consequences	
Physical injury	No physical injuries
	Minor physical injuries
	Major physical injuries
Other damages	Material damage
	Environmental contamination
	Moral prejudice

3.1 Dimension changeability

The dimension changeability describes changes in the characteristics or behavior of technical-physical systems during operation, as well as changes in the environment of the system. With respect to the system in operation, four levels of changeability are distinguished, for which both type and extent of changeability can be very different. These include, among others, "predictable and controllable changes after delivery" of the system, which are either initiated by the user or the manufacturer/operator, or which are designed into the system itself by the latter. Among the four levels of system changeability, "Major changes in operational use" bring up a new quality, as changes and their consequences can no longer be fully predicted. Because the system uses data accumulated during operation to modify its own behavior (usually an optimization), system behavior might change fundamentally over time. Changes during operation on the basis of additional data was referred to in the project as "continued learning". For legal evaluation, systems that continue to learn are particularly significant.

3.2 Dimension transparency

The dimension transparency of the taxonomy describes the comprehensibility of the system or its behavior from two perspectives: 1) An expert's point of view, i.e., the developer or safety engineer, and 2) the completeness of information available for and the comprehensibility of a system to a user, operator, maintainer, or otherwise involved party with respect to the information needs concerning the general and situation-specific functioning. From an expert's point of view, transparency refers to the ability for sprecification, definability of the limits of capability, understandability and predictability of the system's behavior. For other parties involved, knowledge of the scope and limits of the system and its dynamics is important.

3.3 Dimension interconnectedness

The dimension interconnectedness describes system characteristics with respect to digital data exchange. It distinguishes between the influence of internal versus external digital data on system behavior. Neither low-dimensional, digitally mediated sensor data for e.g. temperature, pressure, etc. nor communication or interaction with participants or users for the exchange of information about the status, tasks or intentions of the system are included in this dimension. Internal interconnectedness refers to the exchange of digital data between subsystems. External exchange either involves a central instance that orchestrates the actions of the system (and possibly other systems) following a higher-level objective (e.g., demand-oriented optimization), or it involves a distributed network. Interconnectedness is important for safety considerations because there is a risk of a lack of controllability, especially for distributed networks.

3.4 Dimension controllability

The dimension controllability describes the properties of systems and its surroundings with regard to the ability of the manufacturer or operator to contain system behavior. The controllability of a system depends both on the degree of emergence of a system, i.e. enablement of behavior from within, and on restrictions imposed on the system, its surroundings or the data flow from outside. Systems are considered emergent in the sense of this taxonomy if they exhibit autonomy or self-organization. The sub-category constraints includes measures that help to ensure safe behavior despite high emergence as well as measures that ensure the control of system behavior in the case of data flow from outside. These include, for example, system restrictions by technological means aimed at monitoring subsystems that are difficult to control.

3.5 Dimension resilience

The dimension resilience describes the ability of systems to avoid safety-relevant errors despite disturbances, or to at least mitigate their consequences. The sub-category robustness describes the avoidance of errors, while the avoidance or mitigation of their harmful consequences is subsumed under the sub-category failure handling. A distinction is made between external (security) and internal disturbances. In contrast to robustness, which is aimed at avoiding system errors or safety-relevant attacks from the outside, failure handling describes the potential of systems to reduce the severity of consequences - up to and including complete avoidance of harmful consequences.

3.6 Dimension human involvement

The safety of systems cannot be assessed without considering human influences. On the one hand, the safety of the overall system can be influenced by the people involved in its operational use ("human as agent"), while on the other hand, these people must be protected from possible hazards ("human put at risk"). These two aspects are considered within the framework of taxonomy in the dimension human involvement. With regard to the human as an agent, different aspects are considered in the taxonomy: 1) the safety-influencing effect of humans in dealing with the system, 2) their possible deviations from the intended use of the system and 3) a deliberate damaging agent sabotaging the system.

3.7 Dimension harmful consequences

If we look at people being exposed to possible undesirable consequences of unintended system behavior, they are classified by their role in the system context, which is linked to both necessity and means for their protection from harmful consequences. These harmful consequences form the seventh dimension of the taxonomy. For the purpose of the project, the sub-category physical injury is of major importance in comparison with other damages – be it material damage, environmental contaminations or moral prejudice. If the use of a system cannot lead to human injury, the other six dimensions can be regarded as irrelevant in the context of safety considerations.

4 Challenges in describing AI systems

The findings of the interviews indicate that the safety experts did not always grasp the limits and capabilities of AI and the terminology used, while the AI experts were not familiar with the procedures, ways of thinking and regulatory frameworks in the field of product safety. For this reason, an extensive analysis was conducted to define the most relevant terms and thereby provide a foundation for their later use in both the taxonomy and the legal assessment.

5 Terms and definitions

5.1 Autonomous vs. automated

It was shown that the use of the terms "autonomous" and "autonomous system" is domain-specific. It also proved difficult to distinguish the term "autonomous" from the term "automated". Therefore a general and context-independent definition of "autonomous" was developed in this project. According to this definition, "autonomous" describes being independent or self-sufficient from something and with respect to something. An autonomous system thus describes a system that is independent from something and with respect to something - it acts independently in a defined sense. There is a strong semantic proximity of autonomy and automatization. In both terms the concept of independence is addressed, while the object of indepencence differs. While automation focusses on indepencence from human actions, autonomy can also refer to independence from other machines or software systems.

5.2 Intelligence vs. artificial intelligence

In the course of the project the term "intelligence" in the context of AI was also defined. A common definition from psychology describes "intelligence" as a "very general mental capacity, which - among other things - includes the ability to reason, to plan, to solve problems, to think abstractly, to understand complex ideas, to learn quickly and to learn from experience. It is thereby dinstinguished from mere book knowledge, a narrow academic special aptitude or test-taking experience. Instead, intelligence reflects a broader and deeper capacity to understand our environment, to 'get it,' to 'make sense of things,' or to 'figure out' what to do"[1] . The term "intelligence" was developed and applied to describe differences between people. An analysis of the use of the term in the context of AI, as well as in other non-human contexts, revealed that a second concept of intelligence has since been established that differs from the psychological one. One of the main differences is that intelligence in AI is described by system constituents, i.e. discrete characteristics, and less by measurable properties. In contrast to the field of psychology, there is no metric for "intelligence" that could be used to compare the degree of intelligence of different machines. Also, the operationalization of the human intelligence metrics cannot be meaningfully applied for machines.

[1] The complete definition is given in the full report.

The term "artificial intelligence, AI" is similarly difficult to define. In addition to describing a subject area or a system, the term "artificial intelligence" is also used to refer to a subfield of computer science, which can certainly be described as strongly heterogeneous. For example, it includes Symbolic AI, Machine Learning, Connectionism, and many other areas. A content-based definition is difficult due to the growing number of subfields of "artificial intelligence". For these reasons, AI was defined in the project as: "A heterogeneous subfield of computer science that originally intended to emulate human cognition". Accordingly, an "AI system" is a system that contains components that are related to AI.

6 Use of safety-critical AI in SMEs

The novelty of the knowledge field of AI-related safety is also shown in the fact that, according to the experts, AI has not yet found its way into safety-critical applications. None of the experts considered himself able to make an estimate about the developments in the years to come. A lack of standards and of best-practice examples were unanimously cited as the main obstacles to their introduction. This is particularly important for small and medium-sized enterprises, who depend on procedural approaches more than large-scale enterprises (e.g. automotive industry) in order to safeguard their own behavior. The standardization committees however complain that the first steps in identifying best practices have to be taken by the companies, whereas they build their standardization procedures on these best practice examples.

7 Special features of AI methods (e. g. artificial neural networks)

The automotive industry pioneers the development of highly automated vehicles and their ultimate goal of driverless vehicles. AI-based algorithms are applied in highly automated vehicles, but also, for example, in mobile robots. They are mainly used to evaluate camera images and other sensors in order to recognize the vehicle's environment, including all static and dynamic objects, and then derive the situation-adapted behavior of the vehicle- or robot-steering algorithms. For image recognition deep neural networks are often used. These are a subgroup of artificial neural networks (ANN), which in turn belong to the field of machine learning – a subfield of AI. ANNs are characterized by a very large number of behavior-determining variables, whose values are determined by data. The data that is used for learning determines the system‘s input-output behavior. Due to the large number of input variables, for which a broad variety of value assignment can occur in the real world, and the large number of possible states of the ANN, it is not possible to accurately predict the output to a known input. The development process of a software’s function therefore shifts from programming its functionality to selecting training data sets when using ANN.

7.1 Challenges for safety verification

The behavior and the safety of systems are therefore largely determined by the selection and quality of the training data. This is a new approach which has been left unconsidered in the traditional procedures of safety verification until now. Powerful data-driven algorithms are used because they can generate system behavior that cannot be accomplished with traditional software. A disadvantage of the algorithms is the reduced understandability of system behavior: The question "why" a specific behavior occurs can only be answered to a very limited extent. This characteristic of these particular AI-based algorithms was emphasized by many experts and described as a lack of transparency. Another characteristic of ANNs in complex applications is their relatively low robustness compared to non-data-driven software, which may show in large changes at the output despite small changes at the input. This behavior is not known in classical algorithms (at least not at this extend) and makes system behavior unpredictable to some degree.

For safety verification, however, the predictability and understandability of system behavior plays a major role. Although researchers are currently working intensively on ways to increase understandability, it is becoming apparent that a new process for ensuring the safety of data-generated system behavior may have to be developed. Closely related to predictability and understandability is the process of specifying system properties. Some experts emphasized that powerful data-driven algorithms are used in particular for complex tasks that cannot be solved with classic software (e.g. in highly automated driving). However, due to the complexity of the task the functionality of the software can no longer be specified to the same extent as it would have been possible in conventional product development. When it is not possible to exactly specify a system‘s functionality, the conventional verification procedures involving the stages specification - verification (test) - validation can no longer be applied. However, it is noteworthy that their reduced ability for specification is caused by the complexity of their application. Should not AI-based algorithms be able to solve the same tasks in the future, their ability for specification will also be reduced.

Particularly relevant for the legal considerations is the fact that data-driven system behavior can be tailored to its specific application by extending the data-based training process into its operational use. This process of continued learning is known from the application of AI-based algorithms in language recognition algorithms. However, it is more difficult to verify their safety due to the changeability in their real-world application and the lack of predictability and robustness previously mentioned.

8 Taxonomy and its influence on legal considerations

These abovementioned characteristics of certain AI-based algorithms such as changeability, transparency, ability for specification and predictability, as well as controllability and robustness are key elements of the taxonomy developed in the research project. Further elements have been added via the application field of robotics (autonomous robots, driverless transport systems, collaborative robots) as well as the resulting interactions with humans (human involvement). Today‘s possibilities of subsystems that are internally (cyber-physical systems) or externally connected have also been considered. Especially the safety verification for externally connected systems seems challenging as the system‘s behavior is also influenced by external digital data.

A significant part of the legal considerations in this research project deals with a possible mutability of systems. Mutable systems are characterized by both changeability during operational use and connectivity. According to product safety law, it is the manufacturer‘s responsibility to place machines on the market, to put them into service and to ensure that they comply with the safety and health protection requirements given by product safety law. When being placed on the market or commissioned, all formal and material requirements must be met. However, systems that continue to learn – i.e. systems that use data collected in operational use to modify their behavior – are not considered in product safety law, since changes to the system after commissioning are not taken account of in the risk assessment, which is carried out in the process of commissioning (and not afterwards).

9 Risk assessments of AI systems

Theoretically, the risk assessment looks at the entire life cycle of the product. It takes changes that occur over time into account (e.g. due to wear and tear) and determines the intended type and duration of use as well as deviations from the intended use. This assessment forms the basis for the "safety integration". According to product safety law the manufacturer's obligation is focussed on the designated time of placing on the market or commissioning. Although the risk assessment also takes into account the time thereafter, it is completed at that time. Should something occur which would have prevented the product from being placed on the market if it had been known beforehand, the market surveillance authorities will take the necessary measures to counter the risks posed by the product. In some product areas, this market surveillance is flanked by a product monitoring obligation of the manufacturer, e.g. in the case of consumer products. However, this rule-exception relationship between completed and comprehensive risk assessment on the one hand (rule) and unforeseen defects occurring after market introduction for which the intervention of market surveillance authorities is required (exception) is not valid for certain AI-based systems. With a high degree of changeability in operation and a low degree of both transparency and controllability for the parties involved, unforeseen system behavior after placing on the market seems no longer the exception, but the rule. System behavior is regarded as not predictable and the risks can only be determined and described to a limited extent (if at all) within the framework of the risk assessment at the time of placing on the market or commissioning. This unpredictable system behavior cannot be regarded a residual risk, because it is desired, e.g. in order to always achieve the optimum efficiency during operation or to be able to react to frequently changing surroundings. Dealing with such system behavior during operation should, however, not be the responsibility of the market surveillance authority, but of the manufacturer, who in turn should not react with a warning, recall or withdrawal, as would be the case with conventional product monitoring. Instead, the manufacturer's duty could be extended to a risk assessment on the entire life cycle of certain AI systems.

10 Adaptation of the terms "product" and "entity of machine"

In addition, the definitions of products and of an entity of machines seem no longer valid for highly connected systems. If the system can connect with other systems after being commissioned, this must be taken into account by the manufacturer’s design. However, the subsequent interconnectedness does not necessarily imply that a new product is created. If safety-relevant interconnectedness with other systems or machines takes place spontaneously after market introduction or commissioning, if it is initiated by the system itself and neither the systems involved nor the duration of interconnectedness can be predicted, then the manufacturer can hardly complete his risk assessment. The question remains whether such a safety-relevant innerconnectedness creates a new overall system and if this leads to further obligations for the manufacturer (not to mention the question who should be considered the manufacturer in this case). At the same time, the manufacturer of each product must ensure that safety-relevant interconnectedness only takes place if a certain level of IT-security is guaranteed for all connected systems.

11 New legal terms: "mutable" products and "product support concept"

In the research project, two approaches, which are built upon each other, were formulated as a consequence of the above mentioned problems:

- A definition for "mutable" products
- An obligation for the manufacturer to introduce a "product support concept".

In the proposal "modification of product definition", the product definition is extended by stating that changes to the product do not create a new product as long as the changeability is intended. The aim is to regulate products that are particularly changeable, intransparent and hard to control ("mutable products"), without making the unpredictability of system behavior that follows from these characteristics the sole subject of market surveillance or product monitoring. During the course of drafting this alternative legislation, the term "mutability" is defined as a combination of the taxonomy dimensions of changeability, transparency and controllability. It is important that the functionality "intended" is precisely defined for the product in question.

In addition, a concept for "time-period-based obligations" of manufacturers to guarantee safety of "mutable" products was developed. According to this approach, the manufacturers can "update" their risk assessment even after a highly changeable, hardly controllable and intransparent product has been made available, and can adapt the safety measures in place accordingly.

As a legal consequence of a "mutable product", the manufacturer develops an adapted "product support concept". According to this approach, the manufacturer is obliged to ensure safety over the entire intended product life cycle. The product support concept includes both the collection and evaluation of information during the operation of the product as well as taking appropriate measures. In contrast to conventional product monitoring, this approach is based on both observing and reacting to unforeseen but assumed system behavior - instead of reacting to undetected but undesirable risks. This approach includes regulations that affect the relationship between manufacturer and user and thereby supplements contract law with a regulatory dimension.

12 Challenges of networked products and data quality

In order to deal with potential problems in the legal evaluation of certain external interconnectedness, proposals for a

- product term for connected products and a
- "product safety law for data"

were outlined. In order to prevent the intended safety-relevant interconnectedness of a product with external systems from leading to a new product, the product term in product safety law could be redefined to the effect that intended interconnectedness, that is explicitly assumed by the manufacturer, does not lead to a new product. This would leave the sole responsibility for including the safety-related implications of interconnectivity in the risk assessment with the manufacturer of the original product. A "product safety law for data" could facilitate this.

Data service providers could be required to guarantee and document a certain safety level for safety-relevant data that is made available to connected systems. Manufacturers and users of highly connected AI-based products will be relieved of the detailed testing of data quality which is documented by appropriate certifications. This facilitates the risk assessment for the manufacturer of connected products. The use of certified data in the context of interconnectedness is presupposed, so that a certain data quality can be assumed in the risk assessment. By this means, a regulated market for safety-relevant data can be created, in which legal certainty is ensured by a specific product safety law.

13 The ePerson as a new legal subject?

The legal discussions are complemented by an analysis of the concept of ePerson. In the public debate about "autonomous acting systems", which are "intelligent", and make decisions for "themselves" without human involvement, the necessity of a legal decoupling - at least in questions of liability – due to the decoupling from human influence has been discussed for some years now. The result of these considerations is the concept of an "electronic person", mainly discussed in robotic applications. This concept would amount to a comprehensive reform, in which a new type of legal subject would be placed alongside natural and legal persons.

However, the legal analysis concludes that the concept of ePerson fails to solve the complex problems of liability law in the use of AI-based systems. Instead, its introduction would raise a number of new legal uncertainties, especially with regard to the identity of the ePerson, its practical implementation in law and the imminent imbalances in the distribution of responsibility. It seems more sensible to further develop the law for AI-based systems on a sector-specific basis where solutions are needed most due to the technological progress. With a sector-specific approach, more targeted and appropriate solutions can be developed that can build on existing liability concepts, i.e. the reversal of the burden of proof and strict liability, in other areas of law. Legislators can thereby rely on a wide range of instruments suitable for regulating AI-based technology, whereby many alleged problems can already be solved perfectly with the existing law and cautious adjustments.

Suggested citation

Jürgensohn, Thomas et al.: 2022. Legal framework for the deployment of autonomous and AI systems. Dortmund: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin. baua: Focus.

English translation. German original available: doi:10.21934/baua:fokus20210602

Gender-neutral language is used in this publication. Where this is not possible or would detract from the readability of the text, terms used to refer to persons include all genders.

Imprint | Publisher: Federal Institute for Occupational Safety and Health (BAuA),
Friedrich-Henkel-Weg 1-25, 44149 Dortmund | Telephone: +49 231 9071-2071,
mail to: *info-zentrum@baua.bund.de*, Internet: *www.baua.de*,
Design: R. Grahl | DOI: 10.21934/baua:focus20220304 | March 2022

baua:
Federal Institute for Occupational
Safety and Health