International Journal of
Environmental Research and Public Health

MDPI

Article

Sources of Risk of AI Systems

André Steimers * and Moritz Schneider

Institute for Occupational Safety and Health of the German Social Accident Health Insurance (IFA), 53757 Sankt Augustin, Germany; moritz.schneider@dguv.de
* Correspondence: andre.steimers@dguv.de

Abstract: Artificial intelligence can be used to realise new types of protective devices and assistance systems, so their importance for occupational safety and health is continuously increasing. However, established risk mitigation measures in software development are only partially suitable for applications in AI systems, which only create new sources of risk. Risk management for systems that for systems using AI must therefore be adapted to the new problems. This work objects to contribute hereto by identifying relevant sources of risk for AI systems. For this purpose, the differences between AI systems, especially those based on modern machine learning methods, and classical software were analysed, and the current research fields of trustworthy AI were evaluated. On this basis, a taxonomy could be created that provides an overview of various AI-specific sources of risk. These new sources of risk should be taken into account in the overall risk assessment of a system based on AI technologies, examined for their criticality and managed accordingly at an early stage to prevent a later system failure.

Keywords: artificial intelligence; risk management; occupational safety; protective devices; assistance systems

Citation: Steimers, A.; Schneider, M. Sources of Risk of AI Systems. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3641. https://doi.org/10.3390/ijerph19063641

Academic Editors: Marc Wittlich, Massimo Esposito and Paul B. Tchounwou

Received: 19 January 2022
Accepted: 16 March 2022
Published: 18 March 2022

1. Introduction

Artificial intelligence (AI) methods are mainly used to solve highly complex tasks, such as processing natural language or classifying objects in images. AI methods do not only allow significantly higher levels of automation to be achieved, but they also open up completely new fields of application [1]. The importance of artificial intelligence is constantly increasing due to ongoing research successes and the introduction of new applications based on this technology. Driven by success in the fields of image recognition, natural language processing and self-driving vehicles, in the coming years, the fast-growing market of artificial intelligence (AI) will play an increasingly significant role in occupational safety [2,3].

Today, the term artificial intelligence is mainly used in the context of machine learning, such as decision trees or support vector machines, but also includes a variety of other applications, such as expert systems or knowledge graphs [4]. A significant subcategory of machine learning is deep learning, which deals with the development and application of deep neural networks. These neural networks are optimised and trained for specific tasks, and they can differ fundamentally in terms of their architecture and mode of operation [5]. An example would be the use of convolutional neural networks in the field of image processing [6].

AI systems are engineered systems that build, maintain, and use a knowledge model to conduct a predefined set of tasks for which no algorithmic process is provided to the system.

Thus, by using artificial intelligence, concepts such as learning, planning, perceiving, communicating and cooperating can be applied to technical systems. These capabilities enable entirely new smart systems and applications, which is why artificial intelligence is often seen as the key technology of the future [7].

Protective devices and control systems based on artificial intelligence have already enabled fully automated vehicles and robots to be created [8,9]. Furthermore, they enable accidents to be prevented by assistance systems capable of recognising hazardous situations [10,11].

However, for the benefit of human safety and health, safe and trustworthy artificial intelligence is required. This is because, despite the rapid, positive progression of this technology and the new prospects for occupational safety, the increasing application of this technology will also produce new risks [12]. Even today, we already face an increasing number of accidents in systems that utilise artificial intelligence [13], including various reports on fatal accidents due to AI-related failures in automated vehicles [14,15].

Established measures of risk reduction in the development of software are limited in their ability to mitigate these risks, and existing safety standards are hardly applicable to AI systems as they do not take into account their technical peculiarities [16]. For example, during verification and validation activities in the software life cycle, the influences of different input values in the system are investigated, but these can be relatively easily mapped by boundary value analyses. In the field of artificial intelligence, however, this is difficult due to the extensive and complex possible state space. These applications have to deal with the influence of many different biases [17], some of which are specific to AI systems and are therefore not considered in the required verification and validation activities of existing software safety standards.

For this reason, the development of safe AI systems requires a good understanding of the components of trustworthy artificial intelligence [18,19] as risk management for systems that utilise AI must be carefully adapted to the new problems associated with this technology.

Some research proposes assurance cases to support the quality assurance and certification of AI applications. These must provide assessable and structured arguments to achieve a certain quality standard [20–22].

However, these works lack a detailed list of concrete criteria. We propose to define these criteria based on the sources of risk for AI systems. These can then be analysed and evaluated within the risk assessment to derive appropriate risk mitigation measures; this proves to be necessary.

So, it is essential to identify these new risks and analyse the impact of AI characteristics on the risk management strategy, depending on the nature of the system under consideration and its application context. International standards for the AI field are mostly still under development and usually only address partial aspects such as the explainability [23] or controllability [24] of such systems or that they are not applicable to the field of safety-related systems [25]. Other legislative documents such as the Proposal for an Artificial Intelligence Act of the European Commission [26], by their very nature, only define generic requirements at a very high level from which relevant risk fields must first be derived.

Recent approaches for identifying and structuring specific sources of risk for AI systems have already identified some of these risks [12]. However, they do not yet consider important aspects such as security, nor do they offer a proposal for a structured process model for a complete risk assessment in the AI field or a complete taxonomy of risk sources, which would be necessary for the development of corresponding standards. Similarly, they give only a brief description of the sources of risk, which makes it difficult to gain a basic understanding of the difficulties associated with them.

Furthermore, care must be taken to ensure that all identified sources of risk are designed in such a way that they can be taken into account in the overall risk assessment of a system based on AI technologies, examined for their criticality and managed accordingly at an early stage to prevent a later failure of the system.

This paper addresses the question of how the risk of an AI application can be assessed in order not only to determine its level, but also to be able to effectively reduce it. In Section 2, a proposal for an AI risk management process is defined, which is based on established risk management standards, such as ISO 12100 [27] and ISO 14971 [28].

Section 3 then deals with an analysis and taxonomy of specific sources of AI risk. In particular, the relationship between AI technologies and their characteristics is analysed to raise awareness of the new challenges that this technology brings. Section 4 discusses the results presented and concludes by highlighting their significance.

2. Materials and Methods

In order to establish a strategy for promoting occupational safety and health that takes a particular technology into account, it is useful to look at the interaction of this technology with possible risks arising from it. It should be noted that there are different definitions of risk depending on the field of application. In general, risk is defined as the impact of uncertainty on objectives [29]. An impact can, therefore, be a deviation in a positive or negative direction, as well as in both directions. Accordingly, risks can also have positive effects, when following the definition of ISO 31000 [29].

In the context of safety-related systems, however, only the negative effects are usually considered at the system level, especially those that relate to the health and integrity of people. An example is given in the definition of risk according to the ISO IEC Guide 51 [30], which also serves as a basis for risk management standards, such as ISO 12100 [27] or ISO 14971 [28]. In these documents, risk is defined as the "combination of the probability of occurrence of harm and the severity of that harm".

The discipline of occupational safety and health also uses the definition of ISO IEC Guide 51 [30], which is why this definition of risk is used in this paper. From a societal perspective, however, the more comprehensive definition of risk is also helpful and often useful.

Usually, risk is presented in terms of the cause of the risk, the potential events, their impact, and their probability [27,28]. Defined risk management processes are a common way of dealing with risks. These iterative risk management processes involve risk assessment and risk reduction. Risk assessment identifies sources of harm and evaluates the related risks for the intended use and the reasonably foreseeable misuse of the product or system. Risk reduction reduces risks until they become tolerable. Tolerable risk is a level of risk that is accepted in a given context based on the current state of the art.

In order to set up a risk management process for AI systems in the field of occupational safety and health, it is helpful to familiarise oneself with the standards mentioned above. ISO 12100 [27] is helpful here because it describes the risk management process for machinery. However, since AI systems are usually used for complex tasks in complex environments [31], and particularly deep-learning-based models are highly complex [32], the ISO 12100 [27] process needs to be modified somewhat to take these particularities into account.

The ISO 14971 [28] process is helpful in this regard, as active medical devices interact especially often with complex systems, in this case the human body.

Since this complexity results in certain uncertainties, and the complete testing of such systems with all possible interactions can be ruled out, field studies [33,34] are also used here in addition to other test measures in the verification and validation phase, but market monitoring [28] is also required in the risk assessment, in order to be able to carry out field safety corrective actions for distributed products if necessary. This measure is, therefore, a useful addition to the ISO 12100 [27] process, for example.

A resulting possible risk management process for AI systems that is general but still detailed is presented below:

1. Definition of risk acceptance criteria:
2. Risk assessment:
    - 2.1. Risk identification:
        - 2.1.1. Purpose;
        - 2.1.2. Identification of hazards (e.g., by FTA [35], FMEA [35], FMEDA [36]).
    - 2.2. Risk analysis:

2.2.1. Extent of damage;
2.2.2. Probability of occurrence;
2.2.2.1 Hazard exposure;
2.2.2.2 Occurrence of a hazard event;
2.2.2.3 Possibility of avoiding or limiting the damage.

3. Risk evaluation:
4. Risk control:
   4.1. Analysis of risk governance options;
   4.2. Implementation of the risk control measures;
   4.3. Evaluation of the residual risk;
   4.4. Risk–benefit analysis;
   4.5. Analysis of the risks arising from risk governance measures;
   4.6. Assessment of the acceptability of the overall residual risk.
5. Market observation.

2.1. Definition of Risk Acceptance Criteria

The first step of a risk management process involves defining the risk acceptance criteria. In this step, the residual risk that is still acceptable or tolerable is defined. This residual risk is determined by the following factors, among others [37,38]:

- Extent of damage;
- Benefit;
- Voluntariness;
- Costs;
- Secondary effects;
- Delay of the damaging event.

The level of tolerable residual risk is, therefore, derived to a large extent from a tacit social consensus. An example of this is the use of nuclear energy in Germany. This technology has the potential to cause an enormous amount of damage, but at the same time offers a high level of benefit. The risk–benefit analysis was assessed either positively or negatively in political discourse, depending on which factor was weighted more heavily. A social consensus was found through corresponding parliamentary majorities. The occurrence of actual damaging events ultimately led to a short-term change in society's position on this topic, which was promptly followed by an equivalent shift in thinking in the political field [39,40].

Today, there are still only a few recognisable social positions on artificial intelligence. Many applications are already being used and accepted subconsciously [41], while other applications are slowly gaining ground and are being more intensively discussed [42]. In areas where the application has a direct influence on people or could influence their health, the European market is still hesitant to accept AI technology [43]. The reasons for this are largely due to the lack of (comprehensive) regulatory provisions and the lack of corresponding technical standards to which they could refer. Where such systems are used, they are often misused [44,45]. This is based on a general lack of knowledge about the realistic possibilities of this technology. Public perception is usually determined either by promising utopian or dystopian scenarios. However, both of these notions are due to unrealistic perceptions of this technology, often stemming from the way it is portrayed in various media [46].

Therefore, for the widespread and socially accepted use of these technologies to be possible, it is necessary to create appropriate preconditions, which are closely linked to the definition of the risk acceptance criteria. The first steps in this direction were brought up by the development of normative foundations, i.e., in ISO IEC JTC 1 SC 42 "Artificial Intelligence" or CEN CLC JTC 21 "Artificial Intelligence". Furthermore, the European Commission published a first proposal for a regulation on artificial intelligence [26].

Aside from the elaboration of regulative measures, it is equally helpful to inform the public about realistic application possibilities and the limits of this technology. Finally, the acceptance of the technology in a social context should be monitored.

2.2. Risk Assessment

The risk assessment consists of two elements: risk identification and risk analysis. First of all, the risk identification step defines the exact purpose of the application and its limits (compare to Determination of limits of machinery [27], Intended use in ISO 12100 [27], and identification of characteristics related to the safety of the medical device in ISO 14971 [28]). This step is of great importance in the field of artificial intelligence, but it is also difficult to implement, as this technology is mainly used in highly complex environments [31]. A peculiarity of these environments is that it is often not possible to completely define their boundaries, which in turn results in uncertainties. Therefore, one of the key questions regarding the use of artificial intelligence is the extent to which this technology can be used in a safety-related environment and how these uncertainties can be minimised. The answers to such questions can usually be found in corresponding regulations and standards. However, these do not yet exist and must be developed simultaneously to answering these questions.

In addition, all sources of risk associated with artificial intelligence must be identified. These include new sources of risk that can specifically occur with artificial intelligence methods, such as deep learning, but also classic sources of risk that contain new aspects in connection with the use of AI. These risk sources were investigated and will be presented in the results.

The subsequent risk analysis finally examines the probability of occurrence and the actual hazard exposure for each individual identified risk (compare to Identification of hazards and hazardous situations/Hazard identification and Risk estimation/Estimation of the risk(s) for each hazardous situation in ISO 12100 [27] and ISO 14971 [28]). Since there often remains little experience in the development and the use of applications based on this technology—which, in turn, means that the handling of the associated risks is usually of an unknown quantity—small- and medium-sized companies that have not already addressed the area of Trusted AI need extensive assistance in the long term.

When conducting a risk assessment of a workplace, its risk is essentially determined by the following three factors [27]:

- Hazard exposure;
- Occurrence of a hazardous event;
- Possibility of avoiding or limiting the harm.

Section 3 (the Results section) describes various AI-specific risk factors for consideration in a risk assessment. These can be analysed and assessed in the context of the specific AI system. If an unacceptable risk is identified, appropriate risk reduction measures tailored to the individual sources of risk described can then be defined and applied.

2.3. Risk Evaluation

The risk evaluation is based on the results of the risk assessment, which evaluates the existing or potential risk with regard to the extent of damage and the probability of occurrence on the one hand and its impact on the application on the other. This step is often carried out together with a third party in order to have a neutral and independent opinion on this critical step of the product life cycle [27,28].

2.4. Risk Control

After the first iteration of the previous steps, the preliminary result of the risk assessment is determined. If this shows that the tolerable risk is exceeded, risk control measures must be applied. After analysing the options for risk control, these must finally be implemented, and the residual risk must be reassessed (compare to Risk reduction in ISO 12100 [27] and Risk control in ISO 14971 [28]).

Usually, risk control measures are hierarchically prioritised. For example, ISO IEC Guide 51 [30] establishes a three-level classification:

1. Inherently safe design;
2. Safeguards and protective devices;
3. Information for end users.

In general, an inherently safe design should always be attempted; in cases where this is not possible, safeguards and protective devices can be used. If all of these measures are not possible, information for the end users is mandatory.

The problem is that the transfer of a concept idea or an existing product to a safety-related application is a difficult undertaking that requires a lot of experience. Not only are the existing regulations and standards a hurdle, but the concrete implementation of measures also poses a significant challenge.

Technical measures are based on the four pillars of inherently safe design, safety reserves, safe failure and safety-related protective measures. In the field of artificial intelligence, however, these have some special features that need to be considered [27].

2.4.1. Inherently Safe Design

In machine learning, the quality of the result depends to a large extent on the quality of the training data. If the training data do not cover the full variance of the test data or contains errors, the model will produce an equally erroneous algorithm [47]. If a very complex model is used, it is very difficult to understand the decision-making process of the algorithm, and thus identify faulty parts of the software. Consequently, it is advantageous to choose models with a low level of complexity that can be interpreted by humans and can, therefore, be checked and maintained. In this way, features that do not contribute to a causal relationship with the result, and would, therefore, lead to erroneous results, can be removed manually. A disadvantage of interpretable models, however, is that their simplicity is often accompanied by a lower quality in terms of the probability of a correct result [48–50].

2.4.2. Safety Margins

When we look at a mechanical system, for example, there is a point at which a load leads to the failure of the system. As this point can usually only be determined within a certain tolerance range, these systems are operated far below these limits by introducing a certain safety margin or safety factor.

Such uncertainties can also be identified in machine learning. For example, there is uncertainty about whether the learning dataset completely covers the distribution of the test data or uncertainty regarding the instantiation of the test data. Insofar as this uncertainty is captured, a safety margin or safety limit range can also be defined for an algorithm that sufficiently delimits the areas of a reliable decision from those in which an uncertainty exists. Therefore, models that can algorithmically calculate a measure for the uncertainty of their prediction are to be preferred [51,52].

For classification problems, for example, the distance from the decision boundary can be used, whereby a large distance means an increase in the reliability of a prediction [53]. At the same time, however, it must be noted that this only applies to areas in which a high number of available training data exists, and thus have a high probability density. The reason for this is that, in areas with a low probability density, there is usually little or even no training data available. This leads to the fact that, in these areas, the decision boundary is determined by inductive errors, and thus a high epistemic uncertainty, which means that the distance from this boundary has no significance with regard to the reliability of the prediction [54].

2.4.3. Safe Failure

One of the most important strategies in safety engineering is the principle of safe failure. Again, it is important to have a measure of the uncertainty of the prediction. If this

is relatively high, the system could request further verification by a human. In the case of a collaborative robot, however, this would also mean that the robot arm would first have to assume a safe state [52].

2.4.4. Safety-Related Protective Measures

Safety-related protective measures can be implemented in a variety of ways and cover a broad spectrum, from external protective devices to quality-assuring processes for error minimisation. The development process for software in the safety-related environment is governed by a wide range of regulations and standards. The IEC 61508 series of standards—"Functional safety of safety-related electrical/electronic/programmable electronic systems", Part 3 "Software requirements" [36]—is a good example of this. This standard also contains many methods that can be applied to avoid and reduce systematic errors during software development. This development process is embedded in the Functional Safety Management (FSM) plan, which, among other things, describes the entire lifecycle of the system in Part 1. In addition, there are some software-related requirements in Part 2 of this series of standards that must also be considered. However, to date there are no regulations that clarify the relationship between functional safety and artificial intelligence or describe special measures for AI systems in a safety-related environment. At the international level, since 2020, initial activities have been underway to describe requirements for the use of artificial intelligence in the context of functionally safe systems [55].

2.5. Market Observation

New technologies, products and applications can bring with them new risks that, in the worst case, can be overlooked or underestimated. In order to identify such risks, consider them in the future or be able to remove a product from the market in time to improve it, it is necessary to observe the market and to analyse negative incidents in connection with the respective product. For this purpose, not only is it necessary to collect and review reports from the public media, but the specialist literature must be also consulted. Furthermore, an analysis of corresponding accident data would be useful for prevention purposes [28] (ISO 13485).

3. Results

The overall risk management process describes a procedure that identifies risks, assesses them, and defines measures to control them. The core of this process is, of course, the risk assessment, as it is here that the prevailing risks for the system at hand are identified and analysed. To be able to assess the hazards emanating from a technical system, a precise analysis of the sources of risk associated with this system is required. For example, the ISO 12100 standard, "Safety of machinery—General principles for design—Risk assessment and risk reduction" [27], specifies the principles of risk assessment and risk reduction for the machinery sector. This standard contains general sources of risks to be assessed for the machinery sector, whereas the standard ISO 14971, "Medical devices—Application of risk management to medical devices" [28], describes principles of risk management in the sector of medical devices.

However, the use of new technologies, such as the various methods used in the field of artificial intelligence, brings new specific sources of risk or gives rise to new aspects of existing sources of risk that need to be assessed. It is therefore of great importance to identify these sources of risk so that they can be assessed as part of the risk assessment of an AI system.

To identify these new AI-specific sources of risk, it is necessary to evaluate various sources. First, current research trends are relevant. In the field of artificial intelligence, the success of probabilistic and statistical approaches to machine learning in recent years has been undeniable [56,57], and interest in this area continues to be unabated [58]. For this reason, these methods need to be analysed in detail to obtain a list of risk sources that have an impact on the safety of AI systems. For example, the ongoing scientific discussion

on the topic of XAI (explainable artificial intelligence) shows that this is one of the core problems of deep learning [59–61]. This problem is a direct result of model complexity, which in turn results from the application of artificial intelligence to complex tasks in complex environments [62,63]. In this context, however, the question is not only to what extent a task can be automated, but also to what extent it should be automated [64,65]. For example, Article 22 of the General Data Protection Regulation of the European Union [66] states that "The data subject shall have the right not to be subject to a decision based solely on automated processing . . . ". Overall, it can be said that privacy issues are particularly important when using machine learning methods, as these are based on the collection and processing of large data sets [67–69]. Security aspects can also have an impact on the safety of the system, making it important to assess the integrity of the safety behaviour against intentional inputs such as attacks. These include, for example, known inputs that destroy the integrity of software execution (e.g., buffer overflow) but also specific inputs that cause AI models to compute poor results without causing software-level malfunctions [70–72]. Another problem comes from automated decision-making systems, as they run the risk of being subject to a bias that leads to discriminatory results. For this reason, fairness is a major issue [73–77]. One problem with any new technology is the lack of experience in using it; a system that is proven in use can usually be trusted more, as it is assumed that any weaknesses have been discovered and fixed and that it has proven itself to be functional. This lack of experience can also be a problem when using new AI procedures. On the other hand, it represents an opportunity to make the general complexity of these systems controllable (cf. proven in use acc. IEC 61508 [36]).

After an analysis of the research literature, from which various sources of risk could be derived, studies were first compared with existing research on measures for the quality assurance of AI systems. The work of Houben et al. [78] should be mentioned, which provides a comprehensive study of measures for realising the safety integrity of AI systems. Based on the research of such quality-assurance measures, the fundamental problem areas addressed by these measures, and whether they correspond to the previous collection of identified risk sources, were investigated. Subsequently, existing regulations and research on the certification of AI applications were examined. Here, the work of Mock et al. [79], for example, provides a broad overview of the requirements from the draft of an AI regulation of the European Union [26], some research by the ISO/IEC, and the High-Level Expert Group of the EU Commission [19]. Furthermore, this work provides a direct comparison between these documents, which makes it possible to check whether the risks identified so far address these requirements. This comparison was complemented by the current work of Working Group 3 of the standardisation body ISO/IEC JTC 1 SC 42. As a result of these steps, the technical risk sources identified so far were complete according to this work. It should be noted, however, that these differed in part in their terminology as well as in their content. For example, "safety" was often directly addressed in the work discussed. However, to our understanding, this property or basic requirement is a result of a reliable and robust system, which in turn is a result of requirements from various other sub-items. Furthermore, this work does not consider legal risks but focuses on technical requirements and measures for the realisation of a safe AI system.

Finally, to complete the evaluation of the identified sources of risk, the risks were applied to various fatal accidents in recent years [14,15]. The premise was that a risk assessment based on an investigation of the aforementioned sources of risk would have had to address the technical deficiencies revealed by the follow-up investigation of these accidents. The previously identified sources of risk proved themselves in this analysis; nevertheless, it turned out that one critical source of risk was missing. For example, some accidents were based on weaknesses in the system hardware [80], which should be included as a source of risk in AI systems, especially since AI-specific peculiarities also exist here.

In order to be able to classify the identified sources of risk in a taxonomy, the components of a trustworthy AI, according to the High-Level Expert Group on Artificial Intelligence (AI HLEG) of the European Commission, should be used as a guideline. Ac-

cording to the "Ethics guidelines for trustworthy AI" of the AI HILEG, trustworthy AI comprises three components, entailing that the actors and processes involved in AI systems (including their development, deployment and use) should be:

I. Lawful—complying with all applicable laws and regulations;
II. Ethical—ensuring adherence to ethical principles and values;
III. Robust—both from a technical and social perspective.

As mentioned, this work deals with purely technical sources of risk or those sources of risk that entail a technical implementation and not with legal issues.

On this basis, the taxonomy of different sources of risk that can influence the trustworthiness of an AI system presented in Figure 1 was drawn up.

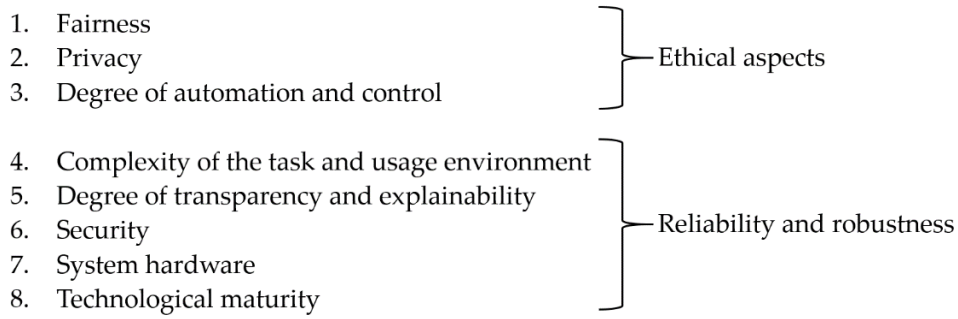

Figure 1. Sources of risk in AI systems that impact the trustworthiness of the system.

These can be roughly divided into two different blocks. The first block deals with ethical aspects. These include fairness, privacy and the degree of automation and control.

The second block deals with various aspects that can influence the reliability and robustness of the AI system, and thus have a direct influence on the safety of the system. Generally, robustness relates to the ability of a system to maintain its level of performance under any circumstances of its usage [4]. Robustness differs from reliability in that a reliable system only needs to maintain its level of performance under the specified conditions for a specific period of time [4]. Robustness, on the other hand, also includes stability against bias or errors and, therefore, represents an extension of the concept of reliability.

In the case of AI, robustness properties demonstrate the ability of the system to maintain the same level of performance when using new data as it achieves when using the data with which it was trained or data for typical operations. Robustness is a new challenge in the context of AI systems, as these systems are used for very complex tasks in complex usage environments, which involve a certain degree of uncertainty. Neural network architectures represent a particularly difficult challenge, as they are both hard to explain and sometimes have unexpected behaviour due to their nonlinear nature. Furthermore, some machine learning methods offer new attack vectors that can reduce the security of the system against external attacks. It is also important to consider the multiple influences of hardware failures, as well as the specific aspects related to them, which can also have a negative effect. Finally, the technological maturity of the AI method used is another important aspect to consider.

Details of the above properties and risk factors, along with their related aspects and challenges, are discussed below.

3.1. Fairness

The general principle of equal treatment requires that an AI system upholds the principle of fairness, both ethically and legally. This means that the same facts are treated equally for each person unless there is an objective justification for unequal treatment.

AI systems used for automated decision-making pose a particular risk for the unfair treatment of specific persons or groups of persons.

To ensure a fair AI system, it must first be investigated whether or to what extent the specific system could make unfair decisions. This depends on various factors, such as the

intended use of the system, as well as the information available for decision-making. If the decisions made by the system cannot have an effect on either natural or legal persons or if the system has no information to distinguish between individuals or groups, it can be assumed that the system does not pose a high risk of discrimination.

The next step is to identify those individuals or groups who can potentially be disadvantaged by the AI system. These can be social minorities or socially disadvantaged groups, but also companies or legal entities in general, as is the case with pricing in digital marketplaces, for example. The General Equal Treatment Act describes various groups of natural persons that are subject to the risk of discrimination:

- Persons of a certain nationality;
- Persons of a certain ethnic origin;
- Persons of a particular gender;
- Persons belonging to a particular religion or belief;
- Persons with a physical or mental disability;
- Persons belonging to a certain age group;
- Persons with a particular sexual identity.

The relevance of these hazards must be assessed in relation to the AI system under investigation. In addition, other groups of people must be considered and added if they could be discriminated against due to the specific context of use or the requirements of the AI application. Likewise, depending on the use case, it must be assessed whether discrimination against legal persons could occur.

If one or more potential hazards to specific individuals or groups are identified, measures must be taken to reduce them to an acceptable level. To do this, it is helpful to look at the causes of possible discrimination.

Unfair outcomes can have several causes, such as bias in objective functions, imbalanced data sets and human biases in training data and in providing feedback to systems. Unfairness might also be caused by a bias issue in the system concept, the problem formulation or choices about when and where to deploy AI systems.

Generally, fairness can be considered as the absence of discrimination in the decisions of an algorithm or in a dataset. For systems whose models were developed on the basis of machine learning methods, the quality of the database is a particularly decisive factor. Here, particular attention must be paid to both the balance and the diversity of the data. Rare cases must not be underrepresented here. Measures that can be used to ensure that datasets are sufficiently balanced and diverse include the massaging, reweighing or sampling of the data.

To prove that an AI system acts fairly, it is necessary to use objective criteria. Fairness metrics are particularly useful here. Such metrics exist for both groups and individuals.

The metrics of group fairness indicate whether all groups of people are treated fairly. When a group is prone to be discriminated against by society, it is referred to as the protected group, whereas a group that is not vulnerable to discrimination is referred to as an unprotected group.

- **Calibration** This states that, for any prediction, samples originating from the protected group must have the same odds to be predicted positively as samples originating from the unprotected group [81].
- **Statistical parity** Statistical parity, also referred to as demographic parity, means that the predictions should be independent of whether a sample belongs to the protected group or not, i.e., the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population as a whole [82–85].
- **Equalised odds** This states that the true positive rates and false positive rates of the outcomes should be the same for the protected group as for the unprotected group. Intuitively, the number of samples classified in the positive class should be the same for the protected group as for the unprotected group [86,87].

- **Equality of opportunity** This is a relaxed version of the equalised odds condition, where it is only required for both the protected and the unprotected group to have an equal true positive rate [88,89].
- **Conditional statistical parity** This states that over a limited set of legitimate factors, the predictions for a sample should be the same for every group. This can be seen as a variant of statistical parity, in which factors that might explain some justified differences in the predictions for two groups are taken into account [85].
- **Predictive equality** This states that decisions made on the protected and unprotected groups should have the same false positive rate. This metric also relates to the equalised odds and equality of opportunity metrics [85]. The metrics of individual fairness focus on individuals and the total set of characteristics that define them, instead of focusing on one characteristic that dictates which group they belong to.
- **Fairness through unawareness** This states that an algorithm is fair if any protected attributes are not explicitly used in the decision-making process. This approach to fairness is also referred to as suppression of the protected attributes [88].
- **Fairness through awareness** This states that an algorithm is fair if similar individuals are treated similarly. By similarity, it is meant that many of their defining characteristics are the same. For every classification task, a similarity measure must be defined. The objective of this measure is to define what exactly is meant by similarity between individuals in the specific context of that task [82].
- **Counterfactual fairness** This states that a decision is fair towards an individual if it coincides with the decision that would have been taken in a counterfactual world. A counterfactual world refers to a situation in which the sensitive attribute is flipped (e.g., flipping the gender). This results in a situation where the individual has the exact same characteristics, apart from the fact that they now belong to the unprotected group instead of the protected one [88].

3.2. Privacy

Privacy is related to the ability of individuals to control or influence what information related to them may be collected and stored and by whom that information may be disclosed. Due to their characteristics, it is possible for AI applications to interfere with a variety of legal positions. Often, these are encroachments on privacy or the right to informational self-determination.

Many AI methods process a variety of different data. Machine learning methods and deep learning methods are especially dependent on large amounts of data, as they need sufficient data to train their models. Ultimately, their accuracy often correlates with the amount of data used. The misuse or disclosure of some data, particularly personal and sensitive data (e.g., health records), could have harmful effects on data subjects. For example, AI applications often process sensitive information, such as personal or private data, including voice recordings, images or videos. Therefore, it must be ensured that the respective local data protection regulations, such as the General Data Protection Regulation (GDPR) [66] in Europe, are observed and complied with.

However, not only can AI applications endanger the privacy of a natural person, but also that of legal persons, for example by releasing trade secrets or license-related data.

Since AI systems often combine data that were previously not linked, it is often possible for them to even capture complex relationships through the creation of extensive models, and thus directly identify persons without even directly specifying the corresponding attributes. However, in addition to the stored or processed data, the ML model implemented in the AI application can also be spied out, which in turn would allow an attacker to extract personal (training) data from a model.

Therefore, privacy protection has become a major concern in Big Data analysis and AI. Considerations regarding whether or not an AI system can infer sensitive personal data should be taken into account. For AI systems, protecting privacy includes protecting the training data used for developing the model, ensuring that the AI system cannot be

used to give unwarranted access to its data, and protecting access to models that have been personalised to an individual or models that can be used to infer information on characteristics of similar individuals.

The risk assessment must determine which specific threats to personal data are posed by the AI system. In particular, the type and significance of the data retrieved or stored during the product lifecycle must be investigated and potential gaps in protection must be identified. It should be noted that this applies not only to data used for development, but also to data used during operation.

The handling of personal data is regulated, for example, by the European General Data Protection Regulation. It should be noted that the legal requirements are not only violated by unauthorised access by third parties, but by the mere existence of unauthorised access, as well as inappropriately long storage periods, and the impossibility to obtain information about the stored data.

Article 5, paragraph 1 of the GDPR [66] describes several principles relating to processing personal data in this regard:

- **Lawfulness, fairness, and transparency** Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject.
- **Purpose limitation** Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.
- **Data minimisation** Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.
- **Accuracy** Personal data shall be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay.
- **Storage limitation** Personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.
- **Integrity and confidentiality** Personal data shall be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction, or damage, using appropriate technical or organisational measures.

If risk control requirements arise, various measures exist to preserve the privacy of personal data. For example, data can be anonymised or pseudonymised. A perturbation or aggregation of data in modelling can also be an effective means of preserving privacy. In general, care should be taken to ensure that only data for a specific purpose are used to the extent necessary.

Another way to prevent unwanted access to data is federated learning, in which several models are trained locally on different computer nodes, so that the respective training data do not have to leave their local position. The separately generated models are then combined into a global model.

3.3. Degree of Automation and Control

The degree of automation and control describes the extent to which an AI system functions independently of human supervision and control.

It thus determines not only how much information about the tactile behaviour of the system is available to the operator, but also defines the control and intervention options of the human. On the one hand, an assessment is made with regard to how high the degree of automation must be for the respective application, but on the other hand, an assessment is also made with regard to whether the human is adequately supported by the AI application and is given appropriate room for manoeuvring in interactions with the AI application. Systems with a high degree of automation may exhibit unexpected behaviour that can be

difficult to detect and control. Highly automated systems can, therefore, pose risks in terms of their reliability and safety.

In this context, several aspects are relevant, such as the responsiveness of the AI system, but also the presence or absence of a critic. In this context, a critic serves to validate or approve automated decisions of the system.

Such a critic can be realised through technical control functions, for example by adding second safety instruments for critical controls that can be understood as an assignment of safety functions to redundant components in the terms of the functional safety standards like IEC 61508-1 [36]. Another way of adding a critic is to use a human whose task is to intervene in critical situations or to acknowledge system decisions. However, even if humans are in the loop and control the actions of a system, this will not automatically reduce such risks and may introduce additional risks due to human variables such as reaction times and understanding of the situation.

Furthermore, the adaptability of the AI system must be considered. Here, the question of whether or to what extent the system can change itself must be considered. Systems that use continuous learning in particular change their behaviour over time. These systems have the advantage of acquiring new functions or adapting to changing environmental conditions via feedback loops or an evaluation function. The disadvantage of such systems, however, is that they can deviate from the initial specification over time and are difficult to validate.

In general, there is a tension between the autonomy of humans and the degree of automation of the AI system. As a rule, a high degree of automation restricts the possibilities of control and influence, and thus, ultimately, the autonomy of humans. It should, therefore, be ensured that human action always takes precedence when using an AI system, i.e., that the human being is always at the centre of the application.

This results in the task of creating an appropriate and responsible distribution of roles between humans and the AI system during the development of such a system. The best way to achieve this is by involving future users as well as domain experts in the development process. The degree of automation must be appropriate to the application context and provide the necessary control options for users. This will ultimately result in a human-centred AI.

In particular, the area of human–machine interaction is the focus for the use of a high degree of automation. AI systems are already being used today in many safety-related applications such as self-driving vehicles [90], aviation [91,92] or the operation of nuclear power plants [93]. In these areas, it is particularly important to ensure that system controls are understandable to people and behave in operation as they would during the design phase.

However, this raises the question of how to manage the uncertainties associated with human–machine interaction with AI-based systems [94]. If human–machine interaction leads to errors or injuries, the question of responsibility arises, as this may be due to incorrect input from the operator but also incorrect or contradictory sensor data. Moreover, in a highly automated system, there is only a limited possibility of human control over the automated system [95,96]. On the other hand, it is also questionable under what conditions an AI system can take control whilst avoiding injuries or errors.

Degrees of automation can be divided into seven different levels, starting from no automation at level 0 to an autonomous system at level 7, which represents the highest level of automation. The SAE standard J3016 [97] defines only six levels for the automotive sector, whereas the standard ISO/IEC 22989 [4] introduces the mentioned seven levels and provides a general description.

It should be noted, however, that today's systems are still all in the area of heteronomous systems, and thus, in practical implementation only a maximum automation level of 6 (full automation) is currently feasible. Table 1 provides an overview and description of the different degrees of automation. The figure also shows how the degree of control by humans decreases as the degree of automation increases.

Table 1. Description of the seven degrees of automation [4]: no automation, assistance, partial automation, conditional automation, high automation, full automation and autonomy.

System	Level of Automation	Degree of Control	Comments
Autonomous	Autonomy	Human out of the loop	The system is capable of modifying its operation domain or its goals without external intervention, control or oversight
Heteronomous	Full automation	Human in the loop Human out of the loop	The system is capable of performing its entire mission without external intervention
	High automation	Human in the loop	The system performs parts of its mission without external intervention
	Conditional automation	Human in the loop	Sustained and specific performance by a system, with an external agent ready to take over when necessary
	Partial automation	Human in the loop	Some sub-functions of the system are fully automated while the system remains under the control of an external agent
	Assistance	Human in the loop	The system assists an operator
	No automation	Human in the loop	The operator fully controls the system

There is some confusion amongst the public, including developers, about the concept of autonomy in the context of AI systems. In general, it must be noted that it is not yet possible to produce artificial autonomous systems by technical means. AI systems as we find them today, can still all be classified as heteronomous systems. Heteronomous systems are distinguished from autonomous systems by being governed by external rules or the fact that they can be controlled by an external agent. In essence, this means that they are operated using rules that are defined or validated by humans. In contrast, an autonomous system is characterised by the fact that it is a system governed by its own rules and not subject to external control or oversight.

A common misconception today is that, in machine learning, the system creates its own rule set, and thus meets the definition for an autonomous system [47,98]. However, it is important to note that these rules are by no means created entirely by the system itself, but rather by the specification of a human-defined algorithm and a training data set determined by the human. Furthermore, these rules are developed to solve a specific task that is also specified by the human. Therefore, in this process, the human not only has complete supervision, but also far-reaching control possibilities.

The concept of autonomy is much broader. In Kant's moral philosophy, for example, the concept of autonomy is defined as the capacity of an agent to act in accordance with objective morality and not under the influence of desires [99]. The concept of autonomy is, therefore, very closely linked to the concept of ethics and, ultimately, to the concept of free will. It is obvious that, to date, there are no AI systems that could be said to have free will, as all AI systems are still completely deterministic systems.

In their work entitled *Moral Machines*, Wallach and Allen [100] specifically address the concept of potentially autonomous machines and view them in the direct context of their ethical sensitivity. They distinguish between operational morality, functional morality and full moral agency. A system has operational morality when the moral significance of its actions are entirely determined by the designers of the system, whereas a system has functional morality when it is able to make moral judgements when choosing an action, without direct human instructions. A fully autonomous system would be a moral agent that has the ability to monitor and regulate its behaviour based on the harm its actions may cause or the duties it may neglect. Thus, a moral actor can not only act morally, but it can also act according to its own moral standards, which means that it would be able to form its own ethical principles or rules. Looking at the capabilities of today's AI systems, only the

level of operational morality can be implemented, so the requirement for an autonomous system is not met.

Wallach and Allan [100] also look at various important abilities that contribute to human decision-making, such as emotions, sociability, semantic understanding and consciousness. As these abilities contribute to human decision-making, they are basic prerequisites for moral systems. If we take only the point of perception here and compare the individual properties associated with human perception with AI systems, it can also be seen that these do not yet fulfil the necessary requirements for autonomous systems. This is summarised again in Table 2.

Table 2. Requirements for autonomous systems and the comparison to current AI systems.

	Autonomous System	**AI System**	
Consciousness			
Memory	Computer memory	No emotional memory	X
Learning	Machine learning	No intuitive learning	X
Anticipation	Predictive analysis	No intuition	X
Awareness	System status	No awareness of self	X
Ethics and morality	Functional morality	No full moral agency	X
Free will	Free decision making	Deterministic systems	X

3.4. Complexity of the Intended Task and Usage Environment

AI is sensibly used for tasks for which there are no classic technologies as alternatives. Such tasks are usually characterised by a high degree of complexity. This complexity can arise from the task itself, as is the case, for example, with the classification of people, or from the complexity of the environment in which it is used, as is the case, for example, in the field of self-driving vehicles. Often, both apply evenly, which makes the task even more difficult.

This complexity gives rise to a certain level of uncertainty in the system's behaviour, which is often perceived as non-deterministic, but whose cause lies in the fact that a complex task and/or environment can only be analysed and described completely by a human with great difficulty. This is mainly due to the large state space of such an environment, which can also be subject to constant change, which, in turn, continuously enlarges the state space, whereby it can be assumed that even a model that generalises the state space very well will not react appropriately to every possible state of the environment.

At this point, however, it should be pointed out once again that AI systems fundamentally work deterministically, even if it may appear otherwise for the reasons mentioned above. The only exceptions are systems that are based on continuous learning and whose models can adapt further during operation.

However, this uncertainty also means that, in the area of safety-related systems, it must be carefully examined whether it is absolutely necessary to use machine learning methods, and in particular deep learning, in the creation of the safety-related system, or whether this could also be carried out using alternative (AI) technologies.

The complexity of the intended task and usage environment of an AI system determines the full range of possible situations that an AI system, when used as intended, must handle. Since it cannot be assumed that it is possible to carry out an accurate and complete analysis of the environment and task to produce the system specification, it will inevitably become relatively vague and incomplete. This may result in the actual operating context deviating from the specified limits during operation.

As a general rule, more complex environments can quickly lead to situations that had not been considered in the design phase of the AI system. Therefore, complex environments can introduce risks with respect to the reliability and safety of an AI system.

For this reason, an AI system must have the ability to still provide reliable results even under small changes in input parameters. Although it is often not possible to predict all

possible states of the environment that an AI system may encounter during its intended use, efforts should be made during the specification phase to gain an understanding of the intended use environment that is as complete as possible. In doing so, knowledge about the input data underlying the decision-making processes and the sources used to obtain it, such as the sensor technology used, should also be obtained and considered. Important aspects here are the questions of whether the system is fed with deterministic or stochastic, episodic or sequential, static or dynamic, and discrete or continuous data.

In the implementation phase, the system is built according to the requirements of the specification. For this purpose, the specification is usually analysed and interpreted by a development team to create a strategy for the technical realization of the system. This strategy and the associated design should again be kept as simple as possible to reduce the complexity of the system and increase its transparency. The implementation strategy is of great importance here, as it can have an immense impact on the design. For example, a requirement can often be paraphrased in such a way that it is still fulfilled semantically, but its technical implementation can be significantly simplified. As an example, consider the simple requirement for a collaborative robot arm that should not reach for a human. This function can already be implemented relatively reliably using deep learning methods, but it is a very complex task for a technical system because the possible state space for the object "human" is very large. If, however, this requirement is formulated in such a way that the robot arm may only grip a certain selection of workpieces, the original goal of the basic requirement is likewise fulfilled but now represents a fairly easy output for the technical system to handle since the object "workpiece" can be completely specified, and thus spans a very small state space.

Another special feature of AI systems based on machine learning methods is the basic implementation process. In classical software development, the specification is interpreted by the development team and implemented accordingly. However, in machine learning systems, which are trained with the help of an algorithm based on data, the mental concept of the specification must be described implicitly by the database. Therefore, the composition of the database and the general data quality are of immense importance. Furthermore, the training algorithm does not always find the best possible solution, which is why it is usually necessary to invest a large amount of work in the optimisation of the resulting model and is standard practice to complete many training runs with different parameterisation, in order to receive a model that is as effective as possible.

Reusing existing components, modules, trained models or complete systems in a new application context can lead to problems due to the different requirements between the specified context and the new context. For example, the use of a system designed to identify people in photos on social networks cannot be easily used to identify people in the context of an assistance system in a work environment. This is partly due to the different state spaces of the applications, so the system may not be able to recognise workers in their personal protective equipment because there are no data to train the model for this but also due to the higher precision required for the latter application. This shows that even a transfer to new environments, for example other industries, is not easily possible.

A few selected examples for model-specific problems regarding the use of trained agents or reinforcement learning are, in no specific order, reward hacking or the safe exploration problem.

The term reward hacking refers to a phenomenon where AI finds a way to gain its reward function, and thus finds a more optimal solution to the proposed problem. This solution, while being more optimal in the mathematical sense, can be dangerous if it violates assumptions and constraints that are present in the intended real-world scenario. For example, an AI system detecting persons based on a camera field might decide that it can achieve very high rewards if it constantly detects persons, and thus will follow them around with its sensors, potentially missing critical events in other affected areas. This can be countered by employing adversarial reward functions, for example, an independent system that can verify the reward claims made by the initial AI and can, most importantly,

learn and adapt. Another option is to pre-train a decoupled reward function that is based solely on the desired outcome and has no direct feedback relation to the initial AI during training.

The safe exploration problem is of particular concern when an agent has the capability to explore and/or manipulate its environment. It is important to note that this does not only pose a problem when talking about service robots, UAVs or other physical entities, but also applies to software agents using reinforcement learning to explore their operating space. In these contexts, exploration is typically rewarded, as this provides the system with new opportunities to learn. While it is obvious that a self-learning AGV needs to follow proper safety protocols when exploring, a system that controls process parameters and employs a random exploration function while not being properly disconnected from the actual process (e.g., via simulation) can pose equal or greater safety risks.

Following proper safety precautions, the first step in ensuring safe operation is typically the application of supervision functions that take over the system in the event that a safety risk is detected, thereby ensuring that no harm can be carried out by the AI. Other options include encoding safe operations as part of the reward function of the system, for example by instructing the model not only to minimise the distance travelled for an AGV, but also to maximise distance to persons. That way, safety precautions become an intrinsic concern of the model by means of the actual reward function.

To cope with the complexity of the task and environment, data quality plays an important role in systems based on machine learning methods. The data must not only be complete, but also diverse, and thus representative enough that a suitable model can be generalised from them. In addition to these two very basic requirements for data quality, there are several other characteristics that must be maintained in order to ensure high data quality:

- **Accuracy** Accuracy is the degree to which data have attributes that correctly reflect the true value of the intended attributes of a concept or event in a particular context of use.
- **Precision** Precision is the extent to which data have attributes that are accurate or allow discrimination in a particular context of use. Precision refers to the closeness of repeated measurements to each other for the same phenomenon, i.e., the extent to which random errors contribute to the measured values.
- **Completeness** Completeness refers to the extent to which a data set contains all of the data it needs to contain.
- **Representativeness** Representativeness refers to the extent to which a data set representing a sample of a larger population has statistical properties that match the properties of the population as defined by the representative sample.
- **Consistency** Consistency refers to the extent to which multiple copies of the same data set contain the same data points with the same values.
- **Relevance** Relevance refers to the extent to which a dataset (assuming it is accurate, complete, consistent, timely, etc.) is appropriate for the task at hand.
- **Data scalability** Data scalability indicates the extent to which data quality is maintained as the volume and velocity of data increases.
- **Context coverage** Context coverage is the degree to which data can be used both in the specified contexts of an ML algorithm and in contexts beyond those originally explicitly specified.
- **Portability** Portability is the degree to which data have attributes that allow them to be installed, replaced or moved from one system to another, while maintaining their existing quality in each context of use.
- **Timeliness** Timeliness indicates the extent to which data from a source arrive quickly enough to be relevant. Timeliness refers to the latency between the time that a phenomenon occurs and the time the data recorded for that phenomenon are available for use; this dimension of data quality is particularly important when the dataset is a continuous stream of data.

- **Currentness** Currentness is the extent to which data have attributes that are the correct age in a particular context of use.
- **Identifiability** Identifiability is the extent to which data can be identified as belonging to a particular person, entity or small group of persons or entities. This concept extends the definition of personal data to entities other than individual persons.
- **Auditability** Auditability refers to the extent to which the quality of the data can be verified.
- **Credibility** Credibility is the degree to which data exhibit attributes that are considered true and believable by users in a particular context of use. Credibility encompasses the concept of authenticity (the truthfulness of origins, attributions, commitments).

Another problem based on the complexity of the task and environment is the potential loss of expressiveness of models. The loss of expressiveness of models is attributed to changes that are historically described by different terms and inconsistently used in the scientific literature. For reference, Moreno-Torres et al. [101] provide an overview of the various terms and their different definitions. In this paper, the two causes of loss of informativeness of models are described by the terms data drift and concept drift.

In data drift, a change in the independent variables (covariates/input characteristics) of the model leads to a change in the joint distribution of the input and output variables. AI components should be inspected for sources of data drift in the context of a safety risk analysis and adequate measures should be planned where necessary. Data drift is often tied to an incomplete representation of the input domain during training. Examples of this include, not accounting for seasonal changes in input data, unforeseen input by operators or the addition of new sensors that become available as input features. Naturally, data drift becomes an issue as soon as a model decays due to a change in the decision boundaries of the model.

Some examples of data drift can be attributed to missing the mark on the best practices in model engineering. Common examples include picking inappropriate training data, i.e., data whose distribution does not reflect the actual distribution encountered in the application context, or even omitting important examples in the training data. As such, these problem instances can be fixed by means of improved modelling and retraining.

Unfortunately, data drift is also caused by external factors, such as seasonal change or a change in process that induces data drift, e.g., replacement of a sensor with a new variant featuring a different bias voltage or encountering different lighting conditions in between training and previously unseen data. It can become necessary for the model to deal with data drift while already deployed, sometimes in cases where retraining is not feasible. In these cases, the model might be constructed in such a way that it is able to estimate correction factors based on features of the input data or allow for supervised correction. Overall, care must be taken to design the model to provide safe outputs, even if there are previously unknown inputs. It is important to understand that, even following proper model engineering practices, such as establishing a sufficiently diverse training dataset, there are no guarantees regarding the resulting model's ability to generalise and adapt to the data encountered in production.

For reference, Amos and Storkey [102] provide illustrations for the most common sources of data drift and provide arguments for model improvements; even when the data drift can be categorised as a simple covariate shift and do not have any apparent effect on classification output, they can lead to simpler or computationally more efficient models. These performance considerations also translate into modern, deep neural networks [103].

Concept drift refers to a change in the relationship between input variables and model output and may be accompanied by a change in the distribution of the input data. Example: the output of a model might be used to gauge the acceptable minimal distance of an operator at runtime based on distance measurements obtained by a time-of-flight sensor (input data). If the accepted safety margins change due to external factors (e.g., increased machine speed not accounted for in the model), concept drift occurs while both processes and inputs have stayed the same.

Systems should incorporate forms of drift detection, distinguish drift from noise present in the system and should ideally adapt to changes over time. Potential approaches include models such as EDDM [104], detecting drift using support vector machines [105], or observing the inference error during training to allow for drift detection and potential adaptation while learning [106]. Furthermore, previous work quantifying drift in machine learning systems is available [107].

Drift is often handled by selecting subsets of the available training data or by assigning weights to individual training instances and then re-training the model. For reference, Gama et al. provide a comprehensive survey of methods that allow a system to deal with drift phenomena [108].

3.5. Degree of Transparency and Explainability

Often, aspects of traceability, explainability, reproducibility and general transparency are summarised under the term "transparency". However, these terms must be clearly distinguished from one another. Transparency is the characteristic of a system that describes the degree to which appropriate information about the system is communicated to relevant stakeholders, whereas explainability describes the property of an AI system to express important factors influencing the results of the AI system in a way that is understandable for humans. For this reason, the transparency of a system is often considered a prerequisite for an explainable AI system. Even if this statement is not entirely correct, in relation to existing model-agnostic methods for increasing the explainability of neural networks, for example, a high degree of transparency nevertheless has a positive effect on the explainability of an AI system.

Information about the model underlying the decision-making process is relevant for transparency. Systems with a low degree of transparency can pose risks in terms of their fairness, security and accountability. Transparency is also a precondition on the reproducibility of the results of the system and bolsters its quality assessment.

The question of whether an AI system is recognisable as such for a user is also answered under this point. On the other hand, a high degree of transparency can lead to confusion due to information overload. It is important to find an appropriate level of transparency to provide developers with opportunities for error identification and correction, and to ensure that a user can trust the AI system.

During a risk assessment, it must be determined which information is relevant for different stakeholders and which risks can result from non-transparent systems for these stakeholders. In this case, a distinction can be made between two main groups of stakeholders:

- **The intended users:** Risks are examined that arise because the decisions and effects of the AI application cannot be adequately explained to users and other affected people.
- **Experts:** Risks are examined that arise because the behaviour of the AI application cannot be sufficiently understood and comprehended by experts such as developers, testers or certifiers.

Table 3 shows some relevant information for different stakeholders. For a developer, the information mentioned under the heading system is particularly relevant, whereas for auditors and certifiers, all the information mentioned is relevant. Users, of course, need to be educated about the nature of the system but only its basic functionality, so the information about the application is particularly interesting for this stakeholder group. However, information such as the objectives of the system and its known constraints are also of high importance for users, as these factors are crucial for safe operation.

Table 3. List of possible information to be communicated to different stakeholders.

System	Data	Application
Design decisions	Place of data collection	Type of application
Assumptions		
Models		
Algorithms		
Training methods	Time of data collection	Degree of automation
Quality assurance processes	Reasons for data collection	Basis of results
Objectives of the system	Scope of data collection	Basis of decisions
Known constraints	Processing of data	User information
	Data protection measures	

With traditional software, the engineer's intentions and knowledge are encoded into the system in a reversible process so that it is possible to trace how and why a particular decision was made by the software. This can be carried out, for example, by backtracking or de-bugging the software. On the other hand, decisions made by AI models, especially by models involving a high level of complexity, are more difficult to understand for humans, as the way knowledge is encoded in the structure of the model and the way decisions are made is rather different from the methods by which humans make decisions. This is especially true for models created with machine learning methods. The methods of deep learning (artificial neural networks) belonging to this category are of particular importance here, as they can sometimes become particularly complex, which means that they are usually almost impossible for a human to explain. Therefore, it is evident that, depending on the type of AI method used, a high degree of transparency does not always automatically lead to a high degree of explainability.

A high level of explainability protects against the unpredictable behaviour of the system but is often accompanied by a lower overall performance in terms of the quality of decisions. Here, a trade-off must often be made between explainability and the performance of a system.

In addition, the accuracy of the information about an AI system's decision-making process is considered in each case. It is possible that a system can provide clear and coherent information about its decision-making process but that this information is inaccurate or incomplete.

Consequently, these aspects should also be included in the general evaluation of the AI system. That way, it is not only examined whether sufficient information about the system is available, but also if it is understandable for both experts and end users, thus delivering reproducible results for users. The question of whether an AI system is recognisable as such for a user is also answered under this point.

The degree of transparency and explainability can be divided into four categories, which are listed below in decreasing order of the degree of transparency and explainability:

1. Explainable:

   The system provides clear and coherent explanations.

2. Articulable:

   The system can extract the most relevant features and roughly represent their interrelationships and interactions.

3. Comprehensible:

   The system is not capable of providing real-time explanations of system behaviour, but these are at least verifiable according to facts.

4. Black Box:

   No information is available about how the system works.

Several evaluation concepts and strategies exist to judge the transparency or even explainability of an AI-based system, such as those reported in [109,110].

Additionally, empirical assessments of the decision process of complex models can be carried out, for example by inspecting a convolutional neural network through the visualisation of the components of its internal layers [111]. The goal is to make the network's decision process more transparent by determining how input features affect the model output. Reviewing the output of a convolutional neural network by having its internal state inspected by a human expert is an approach that is extended in related work, such as [112–114]. Even when access to internal model states is completely unavailable, approaches such as RISE [115] can still provide insights into certain network types.

Even systems traditionally believed to be somewhat explainable with regard to inspection, e.g., decision trees, can quickly reach a complexity that defies understanding when deployed in real-world applications. In situations where an interpretable result is desired, tools, such as optimal classification trees [116] or born-again tree ensembles [117], can be applied to reduce complexity and allow for human expert review.

(See [118] for general thoughts on the relation between AI model types and their interpretability.)

Generally speaking, even if explainable AI is not immediately achievable and might not even be a prime concern when it comes to functional safety, a methodical and formally documented evaluation of model interpretability should be one of the assets employed in safety risk analysis, as this will aid comparability and model selection and can provide insights during a postmortem failure analysis.

3.6. Security

To assess the trustworthiness of an AI-based system, traditional IT security requirements also need to be considered. ISO/IEC 27001 [119], ISO/IEC 18045 [120] and ISO/IEC 62443 [121] already provide processes for the audit and certification of horizontal IT security requirements that are also applicable to AI-based systems.

In addition to following the best practices and observing existing standards for conventional systems, artificial intelligence comes with an intrinsic set of challenges that need to be considered when discussing trustworthiness, especially in the context of functional safety. AI models, especially those with higher complexities (such as neural networks), can exhibit specific weaknesses not found in other types of systems and must, therefore, be subjected to higher levels of scrutiny, especially when deployed in a safety-critical context.

One class of attacks on AI systems in particular has recently garnered interest: adversarial machine learning. Here, an attacker tries to manipulate an AI model to either cause it to malfunction, change the expected model output or obtain information about the model that would otherwise not be available to them.

When trying to manipulate a model, an attacker will typically either modify the input available to the model during inference or try to poison the learning process by injecting malicious data during the training phase. For example, it is possible to trick a model into outputting vastly different results by adding miniscule perturbations to the inputs. This noise is, in the case of input images, generally imperceptible to humans and may also be equally well hidden in numeric inputs. While these perturbations are typically non-random and carefully crafted by means of an optimisation process, it cannot be ruled out that hardware failures or system noise already present in the input can cause a non-negligible shift in model output; see [122], for example. Inputs modified in such a way are called adversarial examples. Adversarial examples translate somewhat well across different model architectures and intrinsic model components [123,124]. This, along with the fact that there are several well-known model architectures and pre-trained models available in so-called "model zoos", makes the practical applicability of adversarial examples seem very likely.

Additionally, even a system that is seemingly resilient to the modification of its inputs, i.e., a system employing a local, non-cloud AI model directly connected to sensors, is not

exempt from this attack vector. The feasibility of physical attacks on models, even if these are considered black boxes with no access to details, the availability of the internal model was already demonstrated by Kurakin et al. in 2017 [125]. More recently, Eykholt et al. [126] showed that it was possible to introduce adversarial examples into the forward inference process of a model by creating the aforementioned perturbations using physical stickers that are applied to objects and cause a vastly diverging classification result. In the examples presented, traffic signs were misclassified with a high success rate [126].

For systems with high demands for safety aspects, these weaknesses should be carefully addressed in terms of both random failures and systematic errors. Overall, failures should be addressed according to best practices, i.e., through hardening, robustification, testing and verification. Additionally, there are specific countermeasures available in the field of machine learning that can be applied to further mitigate the safety risks of AI-specific failure cases. Goodfellow et al. argue that a switch to models employing nonlinear components makes them less susceptible to adversarial examples; however, this comes at the cost of increased computational requirements [127]. Madry et al. [123] address the problem by examining and augmenting the optimisation methods used during training. Often, model ensembles are mentioned to create a more robust overall model through diversification. However, there are results that show that this might not sufficiently harden the system against adversarial examples (see He et al. [128]).

A first step in protecting against attacks on models might be to supply adversarial examples during training, in order to have the model encode knowledge about the expected output of those examples. This is called adversarial training.

The next natural avenue of action involves attempting to remove the artificially introduced perturbations. Some examples of this approach include the high-level representation guided denoiser (HGD) introduced by Liao et al. [129], MagNet, which aims to detect adversarial examples and revert them back to benign data using a reformer network [130] or Defense-GAN, which employs a generative adversarial network with similar goals [131]. It is worth mentioning that scenarios exist where both MagNet and Defense-GAN can fail (see [132]).

Furthermore, noting that the model types typically affected by adversarial attacks are generally robust against noise, several authors propose randomisation schemes to modify the input and increase robustness against malicious, targeted noise. Approaches include random resizing/padding [133], random self-ensembles (RSE) [134] and various input transformations such as JPEG compression or modifications of image bit depth [135]. While these methods can be surprisingly effective, recent results show that these transformations are not sufficient measures under all circumstances. In turn, if input transformations are used as a layer of defence against adversarial examples, the efficiency of said protective measures should be evaluated against examples generated using the Expectation over Transformation (EOT) algorithm presented in [136].

3.7. System Hardware

Of course, an AI model cannot make a course of decisions by itself; it depends on the algorithms, software implementing the AI model and hardware running the AI model. Faults in the hardware can violate the correct execution of any algorithm by violating its control flow. Hardware faults can also cause memory-based errors and interfere with data inputs, such as sensor signals, thereby causing erroneous results, or they can violate the results in a direct way through damaged outputs. This section describes some hardware aspects that can potentially affect the safety of an AI system. As a short summary, currently, we seem to need hardware that is as reliable as the hardware used for conventional systems. In general, hardware-related failures can be divided into three groups:

- Random hardware failures;
- Common cause failures;
- Systematic failures.

Similar to hardware used to execute conventional software, the hardware used to execute AI models also suffers from random hardware failure. These failures include short circuits or interruptions in conductor paths and component parts, short circuits between individual or multiple memory cells of variable and invariable memory, drifting clocks such as oscillators, crystals or PLLs (phase locked loops), and stuck-at errors or parasitic oscillations at the inputs or outputs of integrated circuits. Aside from these failure modes, soft errors can also have an effect. These types of random hardware failures describe unwanted temporary state changes in memory cells or logic components that are usually caused by high-energy radiation from sources such as alpha particles from package decay, neutrons and external EMI (electro-magnetic interference) noise, but can also be caused by internal crosstalk between conductor paths or component parts.

On the downside, when compared to conventional software, computations involving AI models require significantly larger amounts of data movements and arithmetic computations, depending on the types of models used. This may cause a higher probability of faults becoming actual failures, i.e., a higher probability of failure on demand per hour. Furthermore, the training of a model derived from a machine learning method and its execution usually takes place in different systems. Since both faults that occur during the training phase and in the operation of an AI system can affect the correct execution of the algorithm, both the system used for training and the system used for the execution of the AI algorithm are relevant.

In the context of artificial intelligence, GPUs (graphics processing unit), cloud computing or edge computing are the most common methods used for the execution and/or training of the AI algorithm.

Generally, GPUs share their error models with those of a central processing unit (CPU), which means that errors can occur in registers, RAM, address computation, programme counters and stack pointers, for example. The main difference between a CPU and a GPU is the memory and processor architecture. GPUs consist of many multiprocessors that each consist of several processor cores. Each of these multiprocessors is equipped with its own L1 cache, and these caches are not coherent. Compared to the L1 cache of a CPU, the GPU's L1 cache is smaller but has a higher available bandwidth. Unlike the L1 cache, the L2 cache of a GPU is coherent but again smaller than the L2 or L3 cache of a CPU. Because of this, memory diagnostics measures are more challenging to implement on a GPU and an erroneous thread scheduler has a more critical effect.

The cloud computing method is characterised by the fact that it accesses a shared pool of configurable computing resources (such as servers, storage systems, applications and services) over the Internet, whereas edge computing is characterised by local computing power that is in close proximity to the attached devices and provides real-time capability. Since the exchange of data plays a central role in both technologies, it is of particular importance to perform an analysis of the possible faults of the network architectures used. A fault model for different networks includes, for example, errors such as data corruption, unintended repetition of messages, an incorrect sequence of messages, data loss, unacceptable delays, insertion of messages, masquerading and various addressing issues.

On the other hand, there are some reports (for example references [137–139]) suggesting that some internal redundancy of computations embedded in AI models will suppress the negative effects of soft errors to some extent. Despite this, it is difficult to predict the levels of such error suppression with any degree of reliability. The analysis of vulnerability factors for AI is an important aspect of random hardware failure in the context of functional safety.

Common cause failures can be created by AI at the hardware level, as the amount of power required to perform calculations and the loads on system design can vary depending on the data. As AI implementation typically requires more computation resources than the same functionality implemented in conventional software, careful hardware design and implementation is essential. With regard to common cause failures at the hardware level, there are no differences between conventional and AI-based hardware. A list of

relevant common cause failures can be found in standards such as IEC 61508-2 [36] or ISO 26262-11 [140].

Systematic failures are also a cause of error when it comes to hardware systems for creating, maintaining, or running AI models. As the range of AI applications is expanding, embedded systems are also becoming increasingly important. In the training phase, the amount of data and computing power that is required to calculate the coefficients by a machine learning algorithm is very high and prevents the use of an embedded system during this phase. When the training phase is completed, the calculated coefficients are transferred to the target system. This asymmetry of machine learning methods means that much less computing power is required in the application phase; therefore, embedded systems can be suitable for this phase. However, there are some difficulties in implementing the training outcomes on a micro controller unit (MCU), micro processing unit (MPU) or digital signal processor (DSP), as many AI frameworks use Python as the description language, while the control programme of an embedded system is usually in C or C ++. Aside from this, incompatibilities with the read-only memory (ROM) and random access memory (RAM) management of an MCU, MPU or DSP are an additional cause of errors.

The use of parallel computing architectures also increases the risk of time-related programme errors, such as race conditions, deadlocks or heisenbugs.

A deadlock—also called a jam—is a state of processes in which at least two processes are waiting between each other for resources that are allocated to the other process. Thus, the execution of both processes is blocked.

A race condition is a constellation in which the result of an operation depends on the temporal behaviour of certain individual operations. They are a very difficult source of error to detect because the successful completion of the operation depends on chance.

A Heisenbug is a programme error (also called a bug) that is extremely difficult or impossible to reproduce. The defining characteristic of a Heisenbug is the extremely difficult recovery of the framework conditions necessary for the reproduction of the bug. The cause of this type of error is often the use of an analysis tool or debugger, as these can change the temporal framework conditions for the programme flow, and thus prevent the error from occurring. So, you either know the framework conditions without the bug or the bug without the framework conditions, hence the reference to Heisenberg's uncertainty principle.

Some errors cannot be dedicated to a single pitfall, but instead arise from a combination of different ones. Other failures arise from common cause effects that are often not related to failures of single hardware components, but instead to other effects such as electromagnetic interference, temperature effects or decoding errors. Because of this, it is important to be aware of effects that might influence each other.

The following classification scheme of different integrity levels of the hardware is based on IEC 61508-2 [36]:

1. Quantified hardware, SIL 4 capable;
2. Quantified hardware, SIL 3 capable;
3. Quantified hardware, SIL 2 capable;
4. Quantified hardware, SIL 1 capable;
5. Non-quantified hardware proven in field of application;
6. Non-quantified hardware, proven in field of application;
7. Non-quantified hardware, recently released.

3.8. Technological Maturity

The technological maturity level describes how mature and error-free a certain technology is in a certain application context. If new technologies with a lower level of maturity are used in the development of the AI system, they may contain risks that are still unknown or difficult to assess. Mature technologies, on the other hand, usually have a greater variety of empirical data available, which means that risks can be identified and assessed more easily. However, with mature technologies, there is a risk that risk awareness de-

creases over time. Therefore, positive effects depend on continuous risk monitoring and adequate maintenance.

To determine the maturity level of a system, one can, for example, rely on the market's experience with certain technologies or on a systematic analysis of the system's behaviour in operation. Such an analysis is based on evidence of the system's hours of operation in a similar application context, as well as the evaluation of the incidents reported with this system during this time.

The maturity of a technology for implementing an AI system can be classified as follows:

1. Current: The technology is currently supported and in use.
2. Preferred: The technology is already preferred for the implementation of most applications.
3. Limited: The technology is already operational for the implementation of a limited number of applications.
4. Strategic: The technology is likely to be operational only in the medium-to-long term.
5. Emerging: The technology is being researched and tested for possible future use.
6. Out of service: The technology is on the verge of no longer being used.

4. Conclusions

Artificial intelligence is still a very agile field of research that is making great progress. Essentially, we envisage three pillars derived from the rapid progress in the field of artificial intelligence within the last few years: The objective for this was the technical but also economic availability of a high computing power, which made it possible to significantly reduce the training times for deep neural networks [67,141,142]. The second pillar is the availability of large amounts of data, which makes the meaningful training of these deep neural networks possible [47,67–69], and the third pillar is the spread of the open-source idea [143–145]. This has not only made new methods, but also entire training algorithms or even complete models, quickly accessible to a broader public audience, which can thus be easily taken up by other working groups and directly used or further optimised.

All of these factors are still in place, which means that it can be expected that major advances will continue to be made in this field, resulting in the continued rapid market growth in artificial intelligence. Even though this technology has already established itself permanently on the market in some areas, the fields of application of artificial intelligence will be expanded in the future through the realisation of new innovative applications.

However, care must be taken to ensure that a human-centred approach is always adopted in the development of such systems. For this, compliance with basic safety principles is essential and must fulfil all the framework conditions for trustworthy AI.

This requires a precise understanding of the specific aspects of the individual artificial intelligence processes and their impact on the overall quality of the system in general, as well as its safety.

In particular, AI systems based on machine learning present new challenges for the security integrity of the system. Since their models are not developed directly based on the interpretation of a specification by human developers, but are indirectly derived from data, major difficulties exist, especially in creating the specification. Ashmore et al. (Ashmore, 2019) derived the risk source of an incomplete specification from this. Other works also name these and point out the problem of interpretability [146–148]. However, it can be stated here that the problem of incomplete specification is a consequence of the complexity of the task and operational environment of AI systems, which can thus be regarded as the actual original source of risk. Furthermore, the term interpretability is not defined in the basic standard for AI terminology [4], which instead defines the concept of explainability, also being reflected in the scientific discipline of XAI (explainable AI).

Many papers address the effectiveness of assurance cases for the quality assurance of AI systems [149–151]. An assurance case is defined as a reasoned, verifiable artefact that supports the assertion that a set of overlying assertions are satisfied, including a systematic argument with underlying evidence and explicit assumptions to support those assertions [152]. However, these works lack a detailed list of concrete criteria and only

describe a few cases at a time, such as fairness [151], or only structure them on the basis of life-cycle phases according to standards such as the CRISP-DM [153], which means that comprehensive coverage of relevant risk areas cannot be achieved [21]. International standards for the AI field are still under development and usually only address partial aspects, such as the explainability [23] or controllability [24], of these systems, which are not applicable to the field of safety-related systems [25]. Legislative documents, on the other hand, only contain generic requirements that must first be interpreted and concretised [26]. There are only a few studies that deal with the definition and description of concrete sources of risk for AI, and they describe these only superficially and incompletely [12].

Therefore, a comprehensive and easily applicable list of new risks associated with AI systems, which also includes the field of safety-related systems, does not yet exist. Especially in the field of occupational safety and health, it is therefore necessary to identify these new risks and analyse the impact of AI features on risk management strategies, depending on the type of system under consideration and its context of application.

Therefore, this work attempts to provide a comprehensive collection of the relevant sources of risk for AI systems and to classify them in a meaningful taxonomy. The single sources of risk can be divided into risks that relate more to ethical aspects (fairness, privacy, degree of automation and control) and those that influence the reliability and robustness of the AI system (complexity of the intended task and usage environment, transparency and explainability, security, system hardware and technological maturity).

To facilitate the integration of these risk sources into a risk assessment of a system based on AI technologies, a risk management process for AI systems was further proposed and explained. With the help of this process, the individual sources of risk can be easily analysed and evaluated in terms of their criticality in order to define suitable risk mitigation measures at an early stage, which ultimately lead to a reliable and robust system and prevent unsafe failures of the AI system.

The individual sources of risk mentioned were evaluated through various steps. Not only were they compared with the partial results of other work, but requirements for trustworthy AI were also analysed so that it could be deduced from these whether the sources of risk mentioned were factors that influenced them. Finally, it was investigated whether the vulnerabilities derived from various accidents could have been revealed by these sources in advance through a risk assessment.

The description of the individual steps of the proposed risk assessment, as well as the description of the individual risk factors, provides the necessary basic understanding of these factors in order to achieve easy applicability, so that this work can provide guidance and assistance for the risk assessment of AI systems.

Even though an extensive evaluation of the sources of risk mentioned has been carried out, it must be noted that the taxonomy presented cannot claim ultimate and permanent completeness. This is due to the novelty of many AI methods and the high dynamics of research in the field of artificial intelligence. It cannot be ruled out that new incidents, such as accidents, will reveal new critical weaknesses or that new procedures will be developed that bring new aspects with them.

Furthermore, it can be discussed whether the allocation of the individual risk sources in the taxonomy is as valid as presented. It could be noted here that measures to ensure the fairness of a system also increase its robustness. However, we chose to assign fairness to the ethical aspects because at the risk assessment level this is a more ethical issue and only the measures to address possible discrediting outcomes ultimately intersect with the issue of task complexity and environment of use.

In order to ensure the completeness and validity of the taxonomy presented, it is absolutely necessary to continue monitoring research, especially in the field of the development of new methods in the field of AI, but also by constantly monitoring the development of new methods in the field of AI.

Furthermore, this work can only provide a basic understanding of the individual sources of risk. Therefore, it is necessary to use this taxonomy as a basis for further

investigating each identified risk source in depth, to explore its causes and influences on the system and, in particular, to develop suitable measures for risk reduction.

It is also important to discuss the points mentioned above in the context of international standardisation, where there is a lack of requirements and especially of detailed guidance on the risk assessment of safety-related systems. This would be a great gain, especially in the area of testing and certification of such systems. The work presented can provide important input here. However, it can also provide important support for planners, developers, data scientists and other stakeholders.

Author Contributions: Conceptualization, A.S. and M.S.; methodology, A.S.; validation A.S. and M.S.; formal analysis, A.S.; investigation, A.S.; resources, M.S.; data curation, M.S.; writing—original draft preparation, A.S.; writing—review and editing, M.S.; visualization, A.S.; project administration, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Delponte, L. *European Artificial Intelligence Leadership, the Path for an Integrated Vision*; Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament: Belgium, Brussels, 2018.
2. Charlier, R.; Kloppenburg, S. Artificial Intelligence in HR: A No-Brainer. PwC. Available online: http://www.pwc.nl/nl/assets/documents/artificial-intelligence-in-hr-a-no-brainer.pdf (accessed on 10 October 2021).
3. PwC. AI Will Create as Many Jobs as It Displaces by Boosting Economic Growth. PwC. Available online: https://www.pwc.co.uk/press-room/press-releases/AI-will-create-as-many-jobs-as-it-displaces-by-boosting-economic-growth.html (accessed on 7 August 2021).
4. *ISO/IEC DIS 22989*; Artificial Intelligence Concepts and Terminology. International Organization for Standardization; International Electrotechnical Commission: Geneva, Switzerland, 2021.
5. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
6. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaria, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 1–74. [CrossRef]
7. Aggarwal, K.; Mijwil, M.M.; Al-Mistarehi, A.H.; Alomari, S.; Gök, M.; Alaabdin, A.M.Z.; Abdulrhman, S.H. Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning. *Iraqi J. Comput. Sci. Math.* **2022**, *3*, 115–123.
8. Pillai, R.; Sivathanu, B.; Mariani, M.; Rana, N.P.; Yang, B.; Dwivedi, Y.K. Adoption of AI-empowered industrial robots in auto component manufacturing companies. *Prod. Plan. Control* **2021**, 1–17. [CrossRef]
9. Cupek, R.; Drewniak, M.; Fojcik, M.; Kyrkjebø, E.; Lin, J.C.W.; Mrozek, D.; Øvsthus, K.; Ziebinski, A. Autonomous Guided Vehicles for Smart Industries–The State-of-the-Art and Research Challenges. In *Computational Science, ICCS 2020, Lecture Notes in Computer Science*; Krzhizhanovskaya, V.V., Ed.; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12141.
10. Altendorf, Hand Guard. Available online: https://www.altendorf-handguard.com/en/ (accessed on 10 January 2022).
11. Arcure Group. Blaxtair. Available online: https://blaxtair.com/ (accessed on 10 January 2022).
12. Steimers, A.; Bömer, T. Sources of Risk and Design Principles of Trustworthy Artificial Intelligence. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. AI, Product and Service. HCII 2021. Lecture Notes in Computer Science*; Duffy, V.G., Ed.; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12778, pp. 239–251.
13. Gray, S. List of Driveless Vehicle Accidents. *ITGS News*. 2018. Available online: https://www.itgsnews.com/list-of-driverless-vehicle-accidents/ (accessed on 17 March 2022).
14. Pietsch, B. 2 Killed in Driverless Tesla Car Crash, Officials Say. *New York Times*. 2021. Available online: https://www.nytimes.com/2021/04/18/business/tesla-fatal-crash-texas.html (accessed on 10 January 2022).
15. Wakabayashi, D. Self-Driving Uber Car Kills Pedestrian in Arizona. *New York Times*. 2018. Available online: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html (accessed on 10 January 2022).
16. Salay, R.; Czarnecki, K. Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262. *arXiv* **2018**, arXiv:1808.01614.
17. *ISO/IEC TR 24027*; Information Technology—Artificial Intelligence (AI)-Bias in AI Systems and AI Aided Decision Making. International Electrotechnical Commission; International Organization for Standardization: Geneva, Switzerland, 2021.

18. *ISO/IEC TR 24028*; Information Technology—Artificial Intelligence-Overview of Trustworthiness in Artificial Intelligence. International Electrotechnical Commission; International Organization for Standardization: Geneva, Switzerland, 2020.
19. European Commission. *Directorate-General for Communications Networks, Content and Technology, Ethics Guidelines for Trustworthy AI*; European Commission Publications Office: Brussels, Belgium, 2019.
20. Batarseh, F.A.; Freeman, L.; Huang, C.H. A survey on artificial intelligence assurance. *J. Big Data* **2021**, *8*, 1–30. [CrossRef]
21. Kläs, M.; Adler, R.; Jöckel, L.; Groß, J.; Reich, J. Using complementary risk acceptance criteria to structure assurance cases for safety-critical AI components. In Proceedings of the AISaftey 2021 at International Joint Conference on Artifical Intelligence (IJCAI), Montreal, QC, Canada, 19–26 August 2021; Volume 21. Available online: http://ceur-ws.org/Vol-2916/paper_9.pdf (accessed on 10 January 2022).
22. Takeuchi, H.; Akihara, S.; Yamamoto, S. Deriving successful factors for practical AI system development projects using assurance case. In *Joint Conference on Knowledge-Based Software Engineering*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 22–32.
23. *ISO/IEC AWI TS 6254*; Information Technology-Artificial Intelligence-Objectives and Approaches for Explainability of ML Models and AI Systems. International Electrotechnical Commission; International Organization for Standardization: Geneva, Switzerland, 2021.
24. *ISO/IEC AWI TS 8200*; Information Technology-Artificial Intelligence-Controllability of Automated Artificial Intelligence Systems. International Electrotechnical Commission; International Organization for Standardization: Geneva, Switzerland, 2021.
25. *ISO/IEC DIS 23894*; Information Technology-Artificial Intelligence-Risk Management. International Electrotechnical Commission; International Organization for Standardization: Geneva, Switzerland, 2021.
26. European Commission. *Proposal for a Regulation of the European Parliament and the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Act*; European Commission Publications Office: Brussels, Belgium, 2021.
27. *ISO 12100*; Safety of Machinery-General Principles for Design-Risk Assessment and RISK Reduction. International Organization for Standardization: Geneva, Switzerland, 2011.
28. *ISO 14971*; Medical Devices-Application of Risk Management to Medical Devices. International Organization for Standardization: Geneva, Switzerland, 2019.
29. *ISO 31000*; Risk Management-Guidelines. International Organization for Standardization: Geneva, Switzerland, 2018.
30. *ISO/IEC Guide 51*; Safety Aspects-Guidelines for Their Inclusion in Standards. International Organization for Standardization; International Electrotechnical Commission: Geneva, Switzerland, 2014.
31. Forbes. Artificial Intelligence and Machine Learning to Solve Complex Challenges. Available online: https://www.forbes.com/sites/maxartechnologies/2021/02/17/artificial-intelligence-and-machine-learning-to-solve-complex-challenges (accessed on 11 February 2022).
32. Hu, X.; Chu, L.; Pei, J.; Weiqing, L.; Bian, J. Model complexity of deep learning: A survey. *Knowl. Inf. Syst.* **2021**, *63*, 2585–2619. [CrossRef]
33. *ISO 14155*; Clinical Investigation of Medical Devices for Human Subjects-Good Clinical Practice. International Organization for Standardization: Geneva, Switzerland, 2020.
34. *ISO 13485*; Medical Devices-Quality Management Systems-Requirements for Regulatory Purposes. International Organization for Standardization: Geneva, Switzerland, 2016.
35. Cristea, G.; Constantinescu, D.M. A comparative critical study between FMEA and FTA risk analysis methods. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, England, 2018.
36. *IEC 61508*; Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems. International Electrotechnical Commission: Geneva, Switzerland, 2000.
37. Häring, I. Risk Acceptance Criteria. In *Risk Analysis and Management: Engineering Resilience*; Springer: Berlin/Heidelberg, Germany, 2015.
38. Marhavilas, P.K.; Koulouriotis, D.E. Risk-Acceptance Criteria in Occupational Health and Safety Risk-Assessment-The State-of-the-Art through a Systematic Literature Review. *Safety* **2021**, *7*, 77. [CrossRef]
39. Augustine, D.L. *Taking on Technocracy: Nuclear Power in Germany, 1945 to the Present*; Berghahn Books: New York, NY, USA; Oxford, UK, 2018; Volume 24.
40. Wiliarty, S.E. Nuclear power in Germany and France. *Polity* **2013**, *45*, 281–296. [CrossRef]
41. Lee, R.S. *Artificial Intelligence in Daily Life*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–394.
42. Schönberger, D. Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. *Int. J. Law Inf. Technol.* **2019**, *27*, 171–203.
43. Vu, H.T.; Lim, J. Effects of country and individual factors on public acceptance of artificial intelligence and robotics technologies: A multilevel SEM analysis of 28-country survey data. *Behav. Inf. Technol.* **2019**, 1–14. [CrossRef]
44. Javadi, S.A.; Norval, C.; Cloete, R.; Singh, J. Monitoring AI Services for Misuse. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 19–21 May 2021; pp. 597–607.
45. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv* **2018**, arXiv:1802.07228.
46. Avin, S. Exploring artificial intelligence futures. *J. AI Humanit.* **2019**, *2*, 171–193.

47. Strauß, S. From big data to deep learning: A leap towards strong AI or 'intelligentia obscura'? *Big Data Cogn. Comput.* **2018**, *2*, 16. [CrossRef]
48. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI-Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef]
49. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
50. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In Proceedings of the 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 0210–0215.
51. Mohseni, S.; Pitale, M.; Singh, V.; Wang, Z. Practical solutions for machine learning safety in autonomous vehicles. *arXiv* **2019**, arXiv:1912.09630.
52. Varshney, K.R.; Alemzadeh, H. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big Data* **2017**, *5*, 246–255. [CrossRef]
53. Ducoffe, M.; Precioso, F. Adversarial active learning for deep networks: A margin based approach. *arXiv* **2018**, arXiv:1802.09841.
54. Liu, J.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax Weiss, T.; Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Adv. Neural Inf. Processing Syst.* **2020**, *33*, 7498–7512.
55. *ISO/IEC AWI TR 5469*; Artificial Intelligence-Functional Safety and AI Systems. International Organization for Standardization; International Electrotechnical Commission: Geneva, Switzerland, 2020.
56. The AlphaStar team. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. Available online: https://deepmind.com/blog/alphastar-mastering-real-time-strategy-gamestarcraft-ii/ (accessed on 10 January 2022).
57. Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **2018**, *362*, 1140–1144. [CrossRef]
58. Schoenborn, J.M.; Althoff, K.D. Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions. In Proceedings of the ICCBR: 27th International Conference on Case-Based Reasoning, Workshop on XBR: Case-Based Reasoning for the Explanation of Intelligent Systems, Otzenhausen, Germany, 8–12 September 2019; pp. 51–60.
59. Meske, C.; Bunde, E.; Schneider, J.; Gersch, M. Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Inf. Syst. Manag.* **2022**, *39*, 53–63. [CrossRef]
60. Ahmed, I.; Jeon, G.; Piccialli, F. From Artificial Intelligence to eXplainable Artificial Intelligence in Industry 4.0: A survey on What, How, and Where. *IEEE Trans. Ind. Inform.* **2022**. [CrossRef]
61. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *arXiv* **2022**, arXiv:2201.08164.
62. Sultana, T.; Nemati, H.R. Impact of Explainable AI and Task Complexity on Human-Machine Symbiosis. In Proceedings of the AMCIS 2021, Virtual, 9–13 August 2021; p. 1715f.
63. Zhang, Y.; Shi, X.; Zhang, H.; Cao, Y.; Terzija, V. Review on deep learning applications in frequency analysis and control of modern power system. *Int. J. Electr. Power Energy Syst.* **2022**, *136*, 107744. [CrossRef]
64. Cetindamar, D.; Kitto, K.; Wu, M.; Zhang, Y.; Abedin, B.; Knight, S. Explicating AI Literacy of Employees at Digital Workplaces. *IEEE Trans. Eng. Manag.* **2022**, 1–14. [CrossRef]
65. Wijayati, D.T.; Rahman, Z.; Rahman, M.F.W.; Arifah, I.D.C.; Kautsar, A. A study of artificial intelligence on employee performance and work engagement: The moderating role of change leadership. *Int. J. Manpow.* **2022**. [CrossRef]
66. European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2022, General Data Protection Regulation (GDPR), Regulation (EU) 2016/679*; European Commission Publications Office: Brussels, Belgium, 2016.
67. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 843–852.
68. Halevy, A.; Norvig, P.; Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **2009**, *24*, 8–12. [CrossRef]
69. Nandy, A.; Duan, C.; Kulik, H.J. Audacity of huge: Overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Curr. Opin. Chem. Eng.* **2022**, *36*, 100778. [CrossRef]
70. Akhtar, N.; Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* **2018**, *6*, 14410–14430. [CrossRef]
71. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. A survey on adversarial attacks and defences. *CAAI Trans. Intell. Technol.* **2021**, *6*, 25–45. [CrossRef]
72. Michel, A.; Jha, S.K.; Ewetz, R. A survey on the vulnerability of deep neural networks against adversarial attacks. *Prog. Artif. Intell.* **2022**, 1–11. [CrossRef]
73. Colloff, M.F.; Wade, K.A.; Strange, D. Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychol. Sci.* **2016**, *27*, 1227–1239. [CrossRef] [PubMed]

74. Bennett, C.L.; Keyes, O. What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Access. Comput.* **2020**, *125*, 1. [CrossRef]
75. Nugent, S.; Scott-Parker, S. Recruitment AI has a Disability Problem: Anticipating and mitigating unfair automated hiring decisions. *SocArXiv* **2021**. Available online: https://doi.org/10.31235/osf.io/8sxh7 (accessed on 10 January 2022). [CrossRef]
76. Tischbirek, A. Artificial intelligence and discrimination: Discriminating against discriminatory systems. In *Regulating Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 103–121.
77. Heinrichs, B. Discrimination in the age of artificial intelligence. *AI Soc.* **2022**, *37*, 143–154. [CrossRef]
78. Houben, S.; Abrecht, S.; Akila, M.; Bär, A.; Brockherde, F.; Feifel, P.; Fingscheidt, T.; Gannamaneni, S.S.; Ghobadi, S.E.; Hammam, A.; et al. Inspect, understand, overcome: A survey of practical methods for AI safety. *arXiv* **2021**, arXiv:2104.14235.
79. Mock, M.; Schmitz, A.; Adilova, L.; Becker, D.; Cremers, A.B.; Poretschkin, M. Management System Support for Trustworthy Artificial Intelligence. Available online: http://www.iais.fraunhofer.de/ai-management-study (accessed on 20 November 2021).
80. Lambert, F. Understanding the Fatal Tesla Accident on Autopilot and the NHTSA Probe. Available online: http://electrek.co/2016/07/01/understanding-fatal-tesla-accident-autopilot-nhtsa-probe/ (accessed on 10 January 2022).
81. Barocas, S.; Hardt, M.; Narayanan, A. Fairness and Machine Learning. Available online: http://www.fairmlbook.org (accessed on 26 November 2021).
82. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, 8–12 January 2012; pp. 214–226.
83. Johndrow, J.E.; Lum, K. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *Ann. Appl. Stat.* **2019**, *13*, 189–220. [CrossRef]
84. Fish, B.; Kun, J.; Lelkes, Á.D. A confidence-based approach for balancing fairness and accuracy. In Proceedings of the 2016 SIAM International Conference on Data Mining, SDM 2016, Miami, FL, USA, 5–7 May 2016.
85. Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; Huq, A. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017, Halifax, NS, Canada, 13–17 August 2017; pp. 797–806.
86. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef]
87. Agarwal, A.; Beygelzimer, A.; Dudık, M.; Langford, J.; Wallach, H. A reductions approach to fair classification. *arXiv* **2018**, arXiv:1803.02453.
88. Chiappa, S.; Gillam, T. Path-specific counterfactual fairness. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 33.
89. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3315–3323.
90. Weise, E.; Marsh, A. Google Self-Driving van Involved in Crash in Arizona, Driver Injured (Update). Available online: https://phys.org/news/2018-05-waymo-self-driving-car-collision-arizona.html (accessed on 10 January 2022).
91. Clark, N. Report on '09 Air France Crash Site Conflicting Data in Cockpit. *New York Times*. Available online: https://www.nytimes.com/2012/07/06/world/europe/air-france-flight-447-report-cites-confusion-in-cockpit.html (accessed on 10 January 2022).
92. German, K. 2 Years after Being Grounded, the Boeing 737 Max is Flying Again. Available online: https://www.cnet.com/tech/tech-industry/boeing-737-max-8-all-about-the-aircraft-flight-ban-and-investigations/ (accessed on 10 January 2022).
93. Walker, J.S. *Three Mile Island: A Nuclear Crisis in Historical Perspective*; University of California Press: Berkeley, CA, USA, 2004.
94. Howard, J. Artificial intelligence: Implications for the future of work. *Am. J. Ind. Med.* **2019**, *62*, 917–926. [CrossRef] [PubMed]
95. Cummings, M.L. Automation and accountability in decision support system interface design. *J. Technol. Stud.* **2006**. Published Online. Available online: https://dspace.mit.edu/handle/1721.1/90321 (accessed on 10 January 2022). [CrossRef]
96. Sheridan, T.B. *Human Supervisory Control of Automation. Handbook of Human Factors and Ergonomics*, 5th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2021; pp. 736–760. ISBN 9781119636083.
97. SAE International. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*; SAE: Warrendale, PA, USA, 2021.
98. Natale, S.; Ballatore, A. Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence* **2020**, *26*, 3–18. [CrossRef]
99. Kant, I. *Fundamental Principles of the Metaphysics of Morals*; Abbott, T.K., Ed.; Dover Publications: Mineola, NY, USA, 2005.
100. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: Oxford, UK, 2008.
101. Moreno-Torres, J.G.; Raeder, T.; Alaiz-Rodríguez, R.C.; Nitesh, V.; Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.* **2012**, *45*, 521–530. [CrossRef]
102. Storkey, A.J. When training and test sets are different: Characterising learning transfer. In *Dataset Shift in Machine Learning*; MIT Press: Cambridge, MA, USA, 2009; pp. 3–28.
103. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
104. Baena-Garcıa, M.; del Campo-Ávila, J.F.; Raúl, B.A.; Gavalda, R.; Morales-Bueno, R. Early drift detection method. In *Proceedings of the Fourth International Workshop on Knowledge Discovery from Data Streams*; ECML PKDD: Berlin, Germany, 2006; Volume 6, pp. 77–86. Available online: https://www.cs.upc.edu/~{}abifet/EDDM.pdf (accessed on 10 January 2022).
105. Klinkenberg, R.; Joachims, T. *Detecting Concept Drift with Support Vector Machines*; ICML: Stanford, CA, USA, 2000; pp. 487–494.

106. Gama, J.M.; Pedro, C.G.; Rodrigues, P. Learning with Drift Detection. In *Advances in Artificial Intelligence–SBIA 2004*; Bazzan, A.L.C., Labidi, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 286–295.
107. Goldenberg, I.; Webb, G.I. Survey of Distance Measures for Quantifying Concept Drift and Shift in Numeric Data. *Knowl. Inf. Syst.* **2019**, *60*, 591–615. [CrossRef]
108. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.* **2014**, *46*, 1–37. [CrossRef]
109. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
110. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Interpretable machine learning: Definitions, methods, and applications. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef]
111. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *Lecture Notes in Computer Science–ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 818–833.
112. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef] [PubMed]
113. Selvaraju, R.R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; Batra, D. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv* **2016**, arXiv:1610.02391.
114. Stacke, K.; Eilertsen, G.; Unger, J.; Lundstrom, C. A Closer Look at Domain Shift for Deep Learning in Histopathology. *arXiv* **2019**, arXiv:1909.11575.
115. Petsiuk, V.; Das, A.; Saenko, K. RISE: Randomized Input Sampling for Explanation of Black-box Models. *arXiv* **2018**, arXiv:1806.07421.
116. Bertsimas, D.; Dunn, J. Optimal classification trees. *Mach. Learn.* **2017**, *106*, 1039–1082. [CrossRef]
117. Vidal, T.; Pacheco, T.; Schiffer, M. Born-Again Tree Ensembles. *arXiv* **2020**, arXiv:2003.11132.
118. Lipton, Z.C. The Mythos of Model Interpretability. *arXiv* **2017**, arXiv:1606.03490.
119. *ISO/IEC 27001:2013 including Cor 1:2014 and Cor 2:2015*; Information Technology-Security techniques-Information security management systems–Requirements. International Organization for Standardization; International Electrotechnical Commission: Geneva, Switzerland, 2015.
120. *ISO/IEC 18045*; Information Technology-Security Techniques-Methodology for IT Security Evaluation. International Organization for Standardization; International Electrotechnical Commission: Geneva, Switzerland, 2021.
121. *ISO/IEC 62443*; Industrial Communication Networks–Networks and System Security. International Organization for Standardization; International Electrotechnical Commission: Geneva, Switzerland, 2018.
122. Su, J.; Vargas, D.V.; Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 828–841. [CrossRef]
123. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2019**, arXiv:1706.06083.
124. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2014**, arXiv:1312.6199.
125. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. *arXiv* **2017**, arXiv:1607.02533.
126. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Models. *arXiv* **2018**, arXiv:1707.08945.
127. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2015**, arXiv:1412.6572.
128. He, W.; Wei, J.; Chen, X.; Carlini, N.; Song, D. Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong. *arXiv* **2017**, arXiv:1706.04701.
129. Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; Zhu, J. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
130. Meng, D.; Chen, H. Magnet: A two-pronged defense against adversarial examples. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, Dallas, TX, USA, 30 October–3 November 2017; pp. 135–147.
131. Samangouei, P.; Kabkab, M.; Chellappa, R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *arXiv* **2018**, arXiv:1805.06605.
132. Carlini, N.; Wagner, D. MagNet and "Efficient Defenses Against Adversarial Attacks" are Not Robust to Adversarial Examples. *arXiv* **2017**, arXiv:1711.08478.
133. Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A. Mitigating Adversarial Effects Through Randomization. *arXiv* **2018**, arXiv:1711.08478.
134. Liu, X.; Cheng, M.; Zhang, H.; Hsieh, C.J. Towards robust neural networks via random self-ensemble. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 369–385.
135. Guo, C.; Rana, M.; Cisse, M.; van der Maaten, L. Countering Adversarial Images using Input Transformations. *arXiv* **2018**, arXiv:1711.00117.
136. Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing robust adversarial examples. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 284–293.

137. Li, G.; Hari, S.K.S.; Sullivan, M.; Tsai, T.; Pattabiraman, K.; Emer, J.; Keckler, S.W. Understanding error propagation in deep learning neural network (DNN) accelerators and applications. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, Denver, CO, USA, 12–17 November 2017; pp. 1–12. Available online: https://dl.acm.org/doi/10.1145/3126908.3126964 (accessed on 6 October 2021).
138. Wei, X.; Zhang, R.; Liu, Y.; Yue, H.; Tan, J. Evaluating the Soft Error Resilience of Instructions for GPU Applications. In Proceedings of the 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), New York, NY, USA, 1–3 August 2019; pp. 459–464. Available online: https://ieeexplore.ieee.org/document/8919569/ (accessed on 2 October 2021).
139. Ibrahim, Y.; Wang, H.; Liu, J.; Wei, J.; Chen, L.; Rech, P.; Adam, K.; Guo, G. Soft errors in DNN accelerators: A comprehensive review. *Microelectron. Reliab.* **2020**, *115*, 113969. [CrossRef]
140. *ISO 26262*; Road Vehicles-Functional Safety. International Organization for Standardization: Geneva, Switzerland, 2011.
141. Hwang, T. Computational Power and the Social Impact of Artificial Intelligence. 2018. Available online: https://ssrn.com/abstract=3147971 (accessed on 10 January 2022).
142. Thompson, N.C.; Greenewald, K.; Lee, K.; Manso, G.F. The computational limits of deep learning. *arXiv* **2020**, arXiv:2007.05558.
143. Oxford Analytica. China will make rapid progress in autonomous vehicles. *Emerald Expert Brief.* **2018**. Published Online. [CrossRef]
144. Gulley, M.; Biggs, R. Science Fiction to Science Fact: The Rise of the Machines. Available online: https://global.beyondbullsandbears.com/2017/10/26/science-fiction-to-science-fact-the-rise-of-the-machines/ (accessed on 10 January 2022).
145. Rimi, C. How Open Source Is Accelerating Innovation in AI. Available online: https://www.techerati.com/features-hub/opinions/open-source-key-ai-cloud-2019-machine-learning/ (accessed on 10 January 2022).
146. Felderer, M.; Ramler, R. Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session). In *Proceedings of the International Conference on Software Quality, Haikou, China, 6–10 December 2021*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 33–42.
147. Sämann, T.; Schlicht, P.; Hüger, F. Strategy to increase the safety of a DNN-based perception for HAD systems. *arXiv* **2020**, arXiv:2002.08935.
148. Willers, O.; Sudholt, S.; Raafatnia, S.; Abrecht, S. Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception tasks. In *Proceedings of the International Conference on Computer Safety, Reliability, and Security, York, UK, 7–10 September 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 336–350.
149. Adler, R.; Akram, M.N.; Bauer, P.; Feth, P.; Gerber, P.; Jedlitschka, A.; Jöckel, L.; Kläs, M.; Schneider, D. Hardening of Artificial Neural Networks for Use in Safety-Critical Applications-A Mapping Study. *arXiv* **2019**, arXiv:1909.03036.
150. Zenzic-UK Ltd. Zenzic-Safety-Framework-Report-2.0-Final. 2020. Available online: https://zenzic.io/reports-and-resources/safetycase-framework/ (accessed on 2 June 2021).
151. Hauer, M.P.; Adler, R.; Zweig, K. Assuring Fairness of Algorithmic Decision Making. In Proceedings of the 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), Porto de Galinhas, Brazil, 12–16 April 2021; pp. 110–113.
152. *ISO/IEC/IEEE 15026-1*; Systems and Software Engineering-Systems and Software Assurance-Part 1: Concepts and vocabulary. International Organization for Standardization; International Electrotechnical Commission; Institute of Electrical and Electronics Engineers: Geneva, Switzerland, 2019.
153. Studer, S.; Bui, T.B.; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Müller, K.R. Towards CRISP-ML (Q): A machine learning process model with quality assurance methodology. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 392–413. [CrossRef]