Quo Vadis AI?

How can “digital trust” be a driver for AI? What do companies have to do for a successful AI transformation and how can they prepare for future requirements?

Draft EU Legislation (EU AI Act), BSI AIC4, German KI-Normungsroadmap and Co. – where the journey is taking us and how companies can benefit from it

pwc

www.pwc.de/trusted-ai

Contents

1 Digital trust as a transformation driver

Artificial Intelligence (AI) is already almost everywhere we look in our day-to-day lives: Whether as a personal language assistant in our houses, with speech recognition in cell phones and cars, with chatbots for interactions with customers, or even in medicine for assisting with diagnoses. Even today, commercial enterprises are also using the potential of AI in many different ways – and their further potential applications are enormous. AI is a future-oriented technology and is at the heart of every digital transformation project. Estimates[1] have predicted that German gross domestic product (GDP) will grow by 11.3% by 2030 thanks to AI alone. This equates to a total figure of around EUR 430 billion.

Nevertheless, a successful AI transformation cannot be taken for granted. Currently a very large number of AI systems commonly used on the market come from countries such as China or the US, which have already made significant investments in this technology. However, in Germany many companies still remain hesitant. A Bitkom survey in 2021 found that one in four companies already plan to make investments in AI; while AI is the most important future technology for around two thirds. However, the proportion of those companies in Germany that are actually using AI systems currently stands at only 8%.

AI implementation

We need a transformation driver: Digital trust in AI. Because a lack of trust in this technology is one of the key reasons for its lower level of use, as well as factors such as missing expertise, financial means, regulatory uncertainty in its usage and therefore low security of investment for stakeholders.

Regulations and standardizations for AI systems can offer an orientation, point to best practices for the use of the technology and strengthen trust. In this way, digital trust helps to improve the acceptance of Artificial Intelligence. To this end, companies need an action corridor for the use of this important future-oriented technology in order to preserve and develop our economic strength; this includes a definition of actions AI systems can – and are allowed to – perform, to what extent and with what effect, and what quality requirements companies must observe during their development.

European and national initiatives offer useful approaches

Even today, a number of initiatives offer an orientation to companies and can be taken into account during the development and use of AI: At the European level, this is the proposal published by the European Union for a comprehensive, harmonized legal framework for AI – the Artificial Intelligence Act (EU AI Act). At the national level, the Federal Office for Information Security (Bundesamt für Sicherheit in der Informationstechnik, BSI) defined the "AI Cloud Service Compliance Criteria Catalogue (AIC4)". The EU AI Act threatens to impose sanctions for breaches of standards, but also provides important incentives to build trust and thereby improve planning and investment reliability for the use of AI. In this way, technology can become a new factor in our economic strength: It builds the bridge between the industrial "Made in Germany" standard and digital transformation – with the same high quality standard: "Trusted AI made in Germany".

[1] https://www.pwc.de/de/pressemitteilungen/2018/pwc-studie-beziffert-potenzial-kuenstlicher-intelligenz-auf-430-milliarden-euro.html

2 How standardization and regulation impacts the AI transformation

But why does the innovative AI market need standards and regulations in the first place? Don't these tend to have the effect of slowing down innovation? In order to address these questions, it is worth having a look at the recent past: When innovative cloud computing technology became established around the world, for example, it was actually standards and criteria that helped to safeguard the technology's long-term success. External auditors certified major cloud providers such as Microsoft, Amazon Web Services or Google under cloud compliance standards such as SOC (Service Organization Control) or BSI C5 (Cloud Computing Compliance Criteria Catalogue), for instance. For users of these cloud services, proof of such standards is crucial when they are selecting providers, reaching decisions about the use of technology – and also for compliance considerations and evidence. This gives users regulatory certainty – the foundation for innovation in an ecosystem with various different stakeholders.

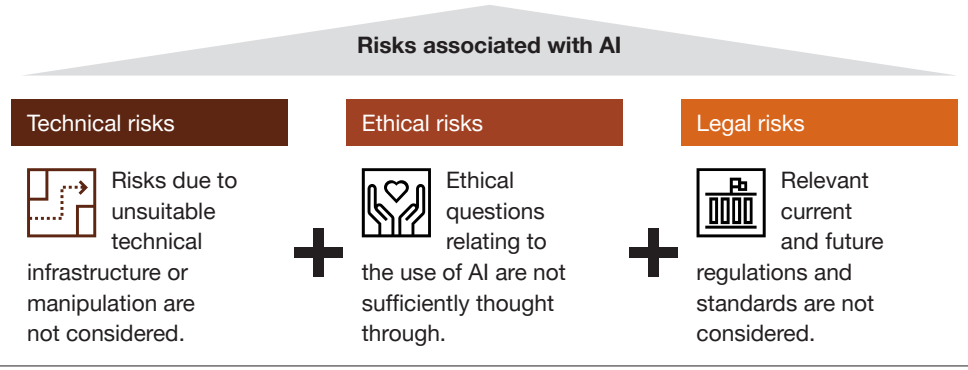
One thing is for sure: New technologies bring new opportunities but also new risks. AI presents various different risks: For example, when deep learning methods are used, the result of an AI system, in other words, the decision it reaches, is not completely transparent or explainable. For some users, AI therefore seems to be something of a "black box". It is all the more important for the quality of an AI system to be presented in a way that is comprehensible. An additional risk is the manipulability of AI. For example, field tests for autonomous driving[2] showed that AI-based driver assistance systems can be deceived relatively easily to intentionally trigger malfunctions if you do not deliberately protect against them.

Moreover: The performance of AI-based algorithms depends to a large extent on the quality of the training data. The value of an AI system therefore highly depends on the value of the data used. Most companies have access to sufficient sensitive data to develop a best-in-class AI system; using available data, however, is often only possible once strict requirements have been met. This is because using this data for a purpose other than that intended may breach privacy rules. And if one trains an AI system with unadjusted data, this can lead to problematic behaviors in the applications so that the algorithm discriminates against individual people or social groups. For example, as part of its proposal to regulate AI the EU therefore plans to prohibit "social scoring", in other words, the technology-based assessment of the population's social behavior. In summary, we can classify AI risks into three groups: technical, legal and ethical risks.

Standards and regulations are effective for these three risk areas because they provide a risk-oriented action framework for the development and use of AI. Uniform standards and criteria build trust among their users and define the quality standards for the use of technology. The example given above of cloud computing showed: In order to achieve scale on the world market, trust is the key to success.

Companies already using AI today have already had to learn this. Because around 70%[3] of today's AI projects do not generate their expected added value. However, this is not due to the technology itself but rather mainly because the relevant stakeholders do not trust the technology. Even today, companies need an organizational and technical framework that provides them with the orientation they need to reliably and securely introduce an AI system. This is the only way they can live up to the "Trusted AI made in Germany" standard.

Different risk groups for artificial intelligence



[2] Yao Deng, Xi Zheng, Tianyi Zhang, Chen, Guannan Lou and Miryung Kim (2020): An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models https://arxiv.org/pdf/2002.02175.pdf
[3] Winning With AI, MITSloan Management Review - By Sam Ransbotham, Shervin Khodabandeh, Ronny Fehling, Burt LaFountain, and David Kiron – https://sloanreview.mit.edu/projects/winning-with-ai/

3 Overview of current standardization and regulation initiatives

But what can companies use for orientation if they want to perform a high-quality AI implementation and build trust in the use of this technology for their customers? What helps are requirements that are developed through policy in conjunction with companies, experts and regulators as part of a public debate. Even today there are various initiatives – from the EU level through to national, industry-specific requirements. All of them seek to strengthen trust in the new technology and drive forward innovation.

3.1 European regulatory initiative: the EU AI Act

With the EU AI Act in April 2021, the European Commission proposed a regulation to develop the European Union into a global hub for trustworthy AI. It seeks to ensure that when AI is used, people's basic rights and the rights of companies are observed in keeping with the EU's ethical values.

At the same time, the EU proposal is intended to promote investment and innovation in secure AI. As with the General Data Protection Regulation (GDPR), the EU consciously opted for the policy instrument of a regulation. This is because this means that the provisions are directly applicable in all member states, which avoids legal fragmentation.

The following are affected by the EU regulation under Article 2 of the EU AI Act:

a) providers placing on the market or putting into service AI systems in the Union, irrespective of whether those providers are established within the Union or in a third country.
b) Users of AI systems located within the Union.
c) providers and users of AI systems that are located in a third country, where the output produced by the system is used in the Union.

The EU AI Act therefore addresses all types of commissioning and distribution of AI systems as well as all types of use within the EU. And this is irrespective of where the provider has its registered office.

So if operators or users of AI systems are located within the EU or if the results obtained from AI systems are used in the EU, the EU AI Act is applicable[4]. It is therefore clear that the regulation applies not only to European actors, but also beyond Europe's borders. However, the same requirements do not apply for every AI system. The intention is to differentiate according to risk categories for which, depending on their classification, more or less strict requirements apply or for which prohibitions may even apply.

1. **AI systems that** do not fall under the following categories 2 to 4 **do not have to meet any requirements.**
2. **AI systems with transparency obligations** such as chatbots. When the system interacts with natural persons you must disclose that they are dealing with an AI system. Such systems should be published in a register.
3. **High-risk AI systems**, for example for biometric identification or AI systems that grant access to general and professional educational institutions. The use of such AI systems also requires that particular requirements and obligations are met under Sections II and III of the EU AI Act, which require that the systems have certain quality and compliance characteristics.
4. **Prohibited AI systems** comprising e.g. subliminal influencing techniques, exploitative practices (the exploitation of weaknesses in a particular group of people or due to age or disability) or social scoring.

Current standardization and regulation initiatives for AI

European proposal for an AI Regulation

Promotion of innovation and investment in secure AI through harmonized regulations at the European level

German AI standardization roadmap ("KI Normung Roadmap")

Specification of an action framework for AI by means of norms and standards, derived from Germany's AI strategy

AI criteria catalog AIC4 published by the BSI

First official criteria catalog for AI services in the cloud environment that defines minimum AI security requirements

[4] The only exception to this, for example, are AI systems used for exclusively military purposes and systems used by public authorities or international organizations that use AI systems in the context of international agreements in the area of criminal prosecutions and judicial cooperation with the Union or with one or several member states.

Should the above requirements be breached, the EU provides for the following sanctions as a deterrent – the higher amount applies in each case:

- For the use or distribution of AI systems belonging to the category of prohibited practices, or in the case of breaches of data and data governance requirements, penalties of up to **EUR 30 million or 6% of annual worldwide turnover** are imposed.
- Further breaches of requirements or obligations such as the absence of a risk management system or the lack of measures to ensure accuracy, robustness and cybersecurity can result in a penalty of up to **EUR 20 million or 4% of annual worldwide turnover**.
- The transmission of false or misleading information to request information from public authorities results in penalties of up to **EUR 10 million or 2% of annual worldwide turnover**.

These sanctions illustrate the grave consequences that companies face if they do not (or cannot) comply with the requirements. Companies should therefore already begin to prepare for the impending regulations now. However, they should not see them as a threat or as hampering innovation. Quite the opposite: The EU measures will help to accelerate innovation. Proactive action to ensure secure, trustworthy AI will promote the long-term success of this technology on the market. It is also useful to take a look at the national initiatives in Germany which, in addition to imposing additional requirements, also first and foremost provide guidance and assistance for implementing the regulation.

3.2 National regulations: German AI Standardization Roadmap and AIC4

The federal government already published its “Artificial Intelligence Strategy” in November 2018. This is intended to provide a policy framework for the holistic continued development and application of AI in Germany. The action areas it describes are explicitly designed to provide a regulatory framework for adherence to the applicable core values and corresponding standards.

To achieve this, the German Institute for Standardization (Deutsche Institut für Normung, DIN) and the German Commission for Electrotechnical, Electronic & Information Technologies (Deutsche Kommission Elektrotechnik Elektronik Informationstechnik, DKE), together with industry experts and the Federal Ministry for Economic Affairs and Energy (Bundesministerium für Wirtschaft und Energie, BMWi) have developed certain norms and standards for AI known as the AI Standardization Roadmap (KI Normungsroadmap). Its goal is to support the international competitiveness of the German industry with the prescribed operational framework.

The participants have defined a total of seven key topics and corresponding operational frameworks here: fundamentals, ethics/ responsible AI, quality, assessment of conformity and certification, IT security for AI systems, industrial automation; mobility and logistics as well as AI in medicine. The implementation of these topics is still ongoing. One means of placing innovative solutions on the market and rapidly setting standards are so-called DIN-SPECs, which can be quickly developed into agile consortia and published within a few months. DIN-SPECs can be the basis for further standardization, the preparation of which allows for the even wider participation of all interested parties. We can assume that additional DIN-SPECs will emerge for dealing with AI.

A catalog of AI criteria sets the benchmark
As mentioned earlier, the first official and specifically auditable requirements for an AI system were provided in February 2021 by the Federal Office for Information Security (Bundesamt für Sicherheit in der Informations-technik) with the criteria catalog for AI services in the cloud environment. AIC4 sets the benchmarks here. This catalog of criteria specifies minimum requirements to safely use machine learning methods in cloud services.

Lifecycle-based approach
AIC4 provides an excellent orientation for the trustworthy handling of AI even for companies that develop and use AI systems outside the cloud. The lifecycle-based, process-oriented approach provides the framework for the development, operation and management of AI. The key areas addressed are Security & Robustness, Performance & Functionality, Reliability, Data Quality, Data Management, Explainability and Bias.

The individual requirements allow an AI system to be evaluated across the entire product lifecycle. They therefore provide a base level of security.

An overview of the 7 BSI AIC4 criteria areas

Security & Robustness
Protection against malicious attacks and interference

Performance & Functionality
Operation of functioning and appropriate AI models

Reliability
Ensuring reliability and bug-fixing measures

Data Quality
Ensuring trust in, and the quality of, used data

Data Management
Ensuring that data is handled properly

Explainability
Ensuring that decisions can be explained

Bias
Avoidance of any unwanted distortion

For cloud services with AI, the AIC4 allows for audits by independent auditors. All companies can use the criteria catalog to derive specific processes, measures and controls for their organization.

In addition to the above initiatives, sector-specific requirements must be observed. For example, AI applications in medicine must meet certain regulatory requirements in order to be admitted to various markets, while in Germany the Medicinal Devices Act (Medizinproduktegesetz) must be observed, among other things. Similarly, AI applications in products used by the automotive industry must comply with the European Regulation on the Registration of Vehicles[5], the requirements of the UNECE (United Nations Economic Commission for Europe) as well as the general requirements of the Product Safety Act (Produktsicherheitsgesetz).

[5] Regulation (EU) 2018/858 of the European Parliament and of the Council of 30 May 2018 on the approval and market surveillance of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles, amending Regulations (EC) No 715/2007 and (EC) No 595/2009 and repealing Directive 2007/46/EC

4 The most important action areas for managing the AI transformation

4.1 AIC4 helps to concretize the requirements of the EU AI Act

The EU AI Act published by the European Commission marks a first step towards encouraging, and demanding, trustworthy AI. When we look at them more closely, however, some of the requirements of this regulation are formulated very concisely in terms of the way that they can be implemented as part of an AI product lifecycle. There is no guidance as to how these requirements can be operationalized.

Below we will describe a selection of requirements that the EU AI Act will impose on providers of high-risk AI systems and for which AIC4 provides specifics and reference points for their operational implementation in processes, controls and measures. A distinction is made between requirements made of the lifecycle process of AI systems and the documentation and information requirements.

Among other things, the EU AI Act requires the establishment of a risk management system and imposes requirements in terms of data and data governance, as well as obligations to keep records, and demands “an appropriate degree of accuracy, robustness and cybersecurity”. These requirements are directly reflected in a similar way that can be operationalized in AIC4.

4.1.1 Risk management requirements

In Article 9 of the EU AI Act on the risk management system the regulation requires that risks emanating from every high-risk AI system be determined on a regular basis. Risk management measures should reduce risks to an acceptable level. Users must regularly check whether these measures are effective.

While the EU AI Act offers little guidance as to how these requirements can be implemented, the AIC4 includes requirements for the management of AI risks, e.g. in the area of Security & Robustness. The underlying methodology and the approach for identifying and evaluating risks is also suitable for preparing for the EU AI Act.

In the case of Security & Robustness, for example, the methodology mentioned above includes screening for possible risk scenarios, the detailed assessment of these scenarios and the countermeasures to be implemented, and testing the suitability and effectiveness of these countermeasures. Both screening and risk scenarios must be regularly checked in order to take account of the latest developments and to evaluate the effectiveness of the countermeasures.

4.1.2 Data management requirements

Article 10 on Data and Data Governance of the EU AI Act imposes quality criteria for training, validation and test datasets, the application of suitable data governance and data management methods as well as the consideration of bias and the protection of personal data. The AIC4 also shows how important data are for AI systems, similarly for training in, and operation of, the systems. This is because it contains separate criteria areas for data management and data quality.

The EU AI Act mentions key aspects – AIC4 performs them in much more detail. The criteria for data quality and management range here from the selection through the development to the operation of the AI system – they therefore cover the entire data lifecycle. It should be noted that data can also come from external data sources, which requires a greater degree of care. The annotation of data is a decisive factor for the quality of the data.

Furthermore, the characteristics and criticality of the AI system must be taken into account in order to adequately address particular features. Aside from the requirements for separating training, validation and test data, it is vital to continually evaluate data quality, both in training and for operations. In this way, countermeasures can then be taken if required in order to safeguard a high level of data quality. For data management, the AIC4 also offers guidance for the administration of access authorizations, the tracking of the data sources and the evaluation of data from external data sources with respect to their credibility and usability.

4.1.3 Documentation requirements

Furthermore, Article 12 of the EU AI Act prescribes record-keeping requirements for the operations and events that occur in AI systems using recognized standards. This allows the activities of an AI system and their triggers to be tracked by automatically logging each operation and its corresponding trigger. This also means that we can ensure – for high-risk AI systems – that the AI system works as intended during its entire lifecycle and that account is taken of critical risk situations.

The EU AI Act does not describe in more detail what situations must explicitly be considered and how an analysis of the log data that are collected should be conducted. The AIC4 shows one operationalization option in the area of reliability: It sets out the requirements for logging, monitoring and the tracking of critical incidents. Additionally, it describes specific details and minimum information for the logs to make it possible to track certain operations. In order to identify abnormalities, the AIC4 also requires the logs to be monitored by means of a connected incident management process.

"Robustness, explainability and transparency are essential to build trust in AI. Governance that is tailored to this sets the pace and is directly reflected in the success of the AI transformation."

– Hendrik Reese

4.1.4 Security requirements

In addition to the risk management system, Article 15 of the EU AI Act requires an appropriate degree of accuracy, robustness and cybersecurity. This includes the provision of indicators for the accuracy of the systems in the instructions for use as well as aspects contributing to the system's resistance to internal errors or external attempts to interfere with it. Technical redundancies or feedback loops are listed as potential measures here.

In the Security & Robustness criteria area, the AIC4 details a methodology for safeguarding the systems' resistance to threats such as external manipulation attempts or internal errors. The methodology corresponds to that used for the risk management system; the perspective changes, however: from the dangers emanating from the AI system towards threats to the AI system. For defining the indicators, the AIC4 also offers an excellent opportunity for operationalization. The Performance & Functionality criteria area describes how indicators can be defined with respect to performance and functionality as well as their monitoring. The AIC4 already takes effect during the development of the AI system by making specific requirements in terms of the selection, training and validation of the AI model being used.

4.1.5 Traceability requirements

In Article 13 on transparency and the provision of information for users, the EU AI Act demands, with respect to the design and development of high-risk AI systems, that their operation is sufficiently transparent and that their results are presented in a way that can be interpreted and easily understood by providers and users. This includes the provision of complete, correct and comprehensible instructions for use. The requirements for this documentation are detailed in the EU AI Act and described in detail. However, the perspective of the technical implementation is missing.

The AIC4 picks up this point in the criteria areas of Explainability and Performance &

Functionality. The catalog provides specific measures with which users can understand and explain the decisions taken by the system. Depending on how sensitive the target application is, the lack of explainability should be made transparent. In this regard, the necessary degree of explainability for the situation in which the AI system is being used should be determined and described.

The parts of the AI system that cannot be explained, technical restrictions to the methods used and inadequacies in relation to the identified requirement for explainability should be taken into account here. Furthermore, for possible effects of bias that can have a critical impact on the system's functionality, metrics and tolerance intervals for evaluating bias that currently cannot be mitigated, must be made clear.

Following the publication of the draft of the EU AI Act, the companies are now in the process of tackling the content of the regulation and implementing it as best as possible. To do this they must perform the corresponding organizational and procedural adjustments in existing process landscapes. Providers must prove – using a conformity evaluation procedure – that they have adhered to the requirements for high-risk AI systems.

Particularly in the areas of Data and Data Governance as well as Accuracy, Robustness and Cybersecurity, the descriptions of the EU Regulation can be interpreted more easily using AIC4 and operationalized using appropriate processes, controls and measures.

4.2 How digital trust can be implemented using the EU AI Act

In order to meet all of the requirements arising from Section II of the EU AI Act, providers of high-risk AI systems must establish reproducible processes and controls. It is therefore essential to build up a solid AI governance system.

The AI applications in companies can be divided into

- AI systems for optimizing internal processes and
- AI systems within services and products that are provided to customers (companies as well as end users).

Depending on the application, a different focus can be used for the implementation, with an AI Governance system providing the framework for both. Particularly for AI use in (end) products, the implementation takes place one level deeper using product compliance management systems.

4.2.1 Implementation of an AI Governance system

Where companies primarily use AI for internal processes, AI-specific aspects can be implemented, orientated according to the AIC4 criteria described above, by extending existing IT governance measures. An AI Governance system with AI-specific guidelines, process and controls ensures the secure and trustworthy handling of the new technology.

AI is also increasingly used in accounting-related processes, with solutions developed in-house, or by external providers. In this way, the AI system can also become the subject of tests during annual audits. The processes and controls that are set up in relation to development, operating and monitoring processes for AI are then the focus of an AI-specific Internal Control System (ICS), which thus becomes increasingly important. Companies should therefore proactively examine their ICS for its applicability and suitability for AI and define additional control measures in a risk-oriented way.

Particularly in order to protect AI against (deliberate or unconscious, e.g. due to insufficient data quality) manipulations, a product-oriented consideration should also occur in addition to the process and control level, as well as including technical validations in the lifecycle management.

4.2.2 Product safety as a goal of AI Governance

For products and services, companies are faced with the question, what measures they should take to securely provide and operate them and develop effective governance systems. In many companies, this task is covered by Compliance Management Systems (CMS). Because the digital transformation not only changes how companies organize themselves internally and what digital tools they use, it also has an impact on their product portfolio.

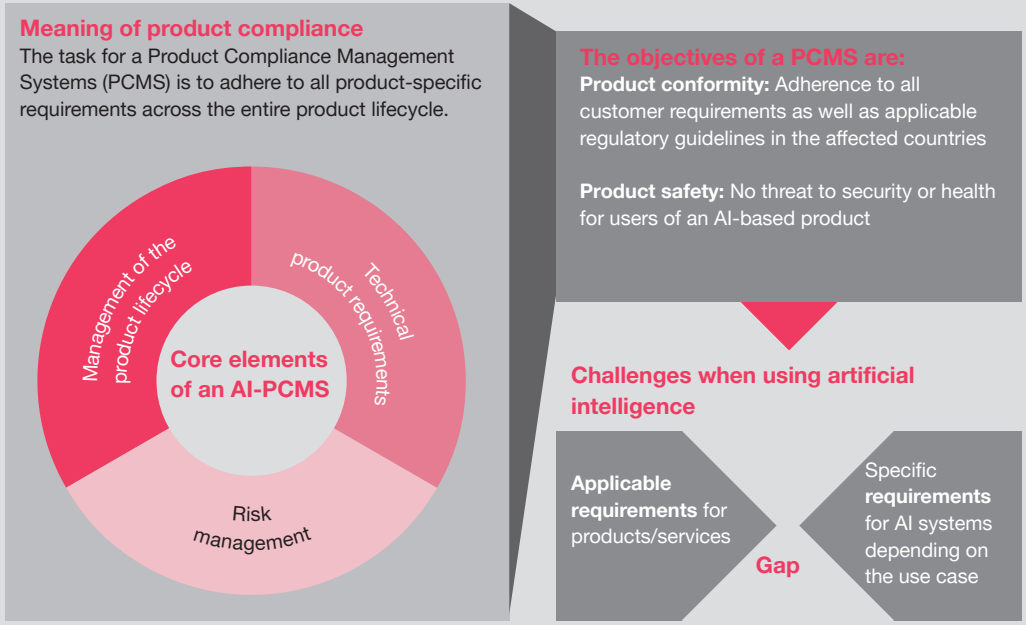
Through internal processes, AI therefore has a direct impact on business partners and end customers. Particularly in order to define measures at the level of products and services, a rethink is required in favor of a Product Compliance Management System (PCMS) that anchors the goals of the AI Governance system in practice.

The term "Product Compliance" can be broken down into product conformity and product safety. What is critical here is that the two elements do not represent a time-based consideration, but rather safeguard a product's compliance over its entire lifecycle. This again underscores the potential of the AIC4 to offer an orientation for practice. Because the AIC4 catalog also conceives the safeguarding of AI as being an integrated task – from development through operation to monitoring.

Product safety – whether with or without AI – is governed in Germany by the Product Safety Act (Produktsicherheitsgesetz, ProdSG), among others. Depending on the application other regulations also play a role, e.g. Regulation 2018/858 for motor vehicle approval. But even the combination of ProdSG and industry-specific regulations is not currently enough to define the technical specifications for AI products or products using AI. The AIC4 can be helpful in the short term because it provides guidance in respect of the arrangement of processes and the development and operation of safe, trustworthy AI and defines minimum requirements in each case.

The achievement of objectives by a CMS largely depends on how well harmonized the management of the product lifecycle is with

Product compliance in the context of Artificial Intelligence



technical product specifications (e.g., derived from the AIC4) as well as the risk management. This is the only way that the needs of different stakeholders can be addressed. Developers of AI systems need e.g. guidelines for the implementation of safety mechanisms that they can translate into software code. Product managers need a risk catalog from which additional risks emerge due to the use of AI compared with conventional IT solutions. And customers need evidence that documents the safety and trustworthiness of AI-based products in the form of a declaration or certificate of conformity.

One example is the UNECE Regulation for the approval of automated lane-keeping systems[6] in motor vehicles. Among other things, it sets out requirements for object recognition[7] by such systems. These also apply to the evaluation of camera images from vehicles that are analyzed by an AI system. In order to ensure that the technology correctly recognizes road geometry, road markings and road signs, specific processes should be defined in the CMS that contribute to this at product level.

The AI system should therefore also be continuously, deliberately trained using manipulated image data. These may be images in which a human figure has been drawn with chalk onto the road surface (risk of emergency braking) or road signs that have been changed using adhesive tape (a speed limit of 130 km/h instead of 30 km/h). By supplying the AI system with these images and flagging them as manipulated data points, the AI can take this into account in its learning process and respond correctly in similar situations.

This illustrates the relevance of an AI-oriented CMS and how it can help to safely make AI-based products ready for the market and safely operate them across their entire lifecycle.

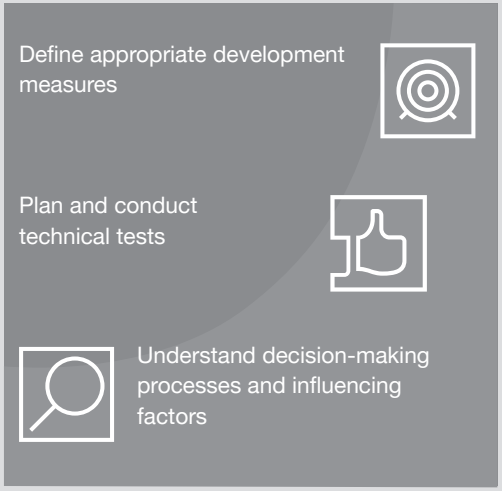
4.2.3 Technical validation

In addition to the introduction of AI-specific processes and controls it may be necessary to perform a technical validation. Rarely does the data already exist in the real world for every possible scenario to allow an AI system to prepare for all eventualities. However, the very strength of AI systems is their ability to reach decisions in unfamiliar situations based on historical data. At the same time, developers and users want to ensure that the AI system is operating within a sensible and safe decision-making corridor. There is a danger, for example, that AI systems may be subject to what is known as 'concept drift' whereby the statistical attributes of the target variables that a model attempts to predict unexpectedly change over time. This means that the system will no longer be optimally used.

Depending on the application area of the AI system it can therefore be very important for companies to understand a model's decision-making processes and influencing factors. Sometimes even minor changes to the input can have a major impact on the processing by the algorithm and therefore the output (e.g. in the case of deliberate manipulations, so-called 'adversarial input'). The technical validation of the model can provide clarity here in order to simulate specific, critical scenarios or examine the response of the model if the input changes.

Technical tests can be used to examine the operation, robustness and transparency of AI systems and, based on the insights obtained, define suitable measures for further development. When developing self-driving cars, for example, it is important to understand why the AI system made a particular – potentially incorrect – decision in a dangerous situation, in order to be able to take specific countermeasures. But how is such a technical validation run?

Undertaking a technical validation



First, it is important to understand what factors can influence the decision of an AI system and to examine which of these factors led to a particular decision. Mathematical methods can be used to explain AI ("explainable AI") in order to assess what contributed to a particular prediction.

[6] https://unece.org/sites/default/files/2021-03/R157e.pdf
[7] Section 7.1 of the UNECE Regulation on the approval of automated lane-keeping systems

One example are applications where an AI system reaches decisions in relation to applications through its output. In these cases it is useful to integrate a fairness principle into the model. There are several definitions for fairness depending on the context, and they contradict each other to some extent. This is why it is important to choose the right fairness metric based on the context.

Depending on this, various evaluation metrics are used to determine whether and to what extent the model is making “unfair” decisions. For example, in order not to disadvantage loan applications on the basis of sex during a loan decision, the “equality of opportunity” metric can be used. This allows the influence of sex on the system’s decision to be evaluated. In this way, developers and users can identify an AI system’s unintentional decision-making factors – thereby immediately giving them a means with which they can put the disadvantaged group on the same footing as others.

From the company’s perspective, this type of technical validation immediately offers a number of advantages: If the available results are explainable and understandable, this will increase the level of trust in the application. This is extremely important in sensitive applications such as in the health or finance sector, but also for self-driving vehicles.

5 Conclusion & outlook: Standards and regulations create trust

The EU AI Act at the European level, and the AIC4 criteria catalog at the national level, got the ball rolling for the standardization and regulation of the responsible use of AI. Standards and regulations are the foundation for building trust in technology for companies and the population and designing future-proof innovation and investments.

Nevertheless, these publications merely represent a first step in the right direction in order to build trust and acceptance within companies, particularly in small and medium-sized companies, as well as among the users of AI systems. They can be regarded as an orientation framework. Nevertheless, it is up to companies themselves how they successfully implement the requirements. They now face the challenge of operationalizing the regulations and standardizations in their companies by embedding them in their processes and internal guidelines.

The longer companies wait for positioning themselves to meet these standards and regulations, the greater the leap that will be required in the future to adjust existing processes. Now is therefore the right time for them to prepare their processes for the existing and future regulations and standardizations and develop short-, medium- and long-term action areas from them. What is clear, however, is that far-reaching, practicable solutions can already be implemented today.

Companies that position themselves as a trustworthy partner for the use of AI on the market at an early stage can establish themselves as pioneers in the trustworthy management of AI, thereby giving them a significant competitive advantage.

However, it is not just companies that have to respond. Standards and regulations typically develop reactively, in other words, with a time delay in response to developments on the market. A strong interplay is required between policy, regulators and industry in order to ensure that standards and regulations continuously reflect the latest developments and the progress of new technology. To this end, it is necessary to continuously develop provisions so that they are always up to date with technological progress. If they succeed in doing so, actors from industry, policymaking and regulation can optimally support the AI transformation.

Points of contact

Hendrik Reese
Director, Trust in AI
Tel: +49 89 5790-6093
E-mail: hendrik.reese@pwc.com

Alexa von Witzleben
Manager, Wirtschaftsprüferin (Auditor), Audit of AI
Tel: +49 89 5790-5487
E-mail: alexa.von.witzleben@pwc.com

Dr Kai Kümmel
Manager, AI Compliance
Tel: +49 89 5790-7153
E-mail: kai.kuemmel@pwc.com

Sources – Introduction:
https://www.bmwi.de/Redaktion/DE/Artikel/Technologie/kuenstliche-intelligenz-warum-deutschland-jetzt-durchstarten-muss-und-kann.html
https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/potenziale-kuenstlichen-intelligenz-im-produzierenden-gewerbe-in-deutschland.html
https://www.bitkom.org/Presse/Presseinformation/Kuenstliche-Intelligenz-kommt-in-Unternehmen-allmaehlich-voran

Sources – Regulation:
https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/fahrplan-festlegen
https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/AIC4/aic4.html
https://www.vde.com/topics-de/health/aktuelles/regulatorische-anforderungen-ki-medizin

www.pwc.de