Artificial intelligence as a driver for economically relevant ecosystems

Technology programme of the Federal Ministry for Economic Affairs and Climate Action

EXPLAINABLE AI

Requirements, Use Cases and Solutions

Study commissioned by the Federal Ministry for Economic Affairs and Climate Action (BMWK) within the framework of the mandated accompanying research for the technology program “Artificial Intelligence as a Driver for Economically Relevant Ecosystems” (AI innovation competition)

www.ki-innovationen.de

IMPRINT

The original study “Explainable AI - Requirements, Use Cases and Solutions” was conducted by the mandated accompanying research for the AI Innovation Competition on behalf of the Federal Ministry for Economic Affairs and Climate Action and published in April 2021. This English translation was published in April 2022.

Publisher

Technology Programme AI Innovation Competition
of the Federal Ministry for Economic Affairs and Climate Action
Accompanying research
iit - Institute for Innovation and Technology in the VDI/VDE Innovation + Technik GmbH
Dr. Steffen Wischmann
Steinplatz 1
10623 Berlin
wischmann@iit-berlin.de

Authors

Dr. Tom Kraus
Lene Ganschow
Marlene Eisenträger
Dr. Steffen Wischmann

Design

LHLK Agentur für Kommunikation GmbH
Hauptstr. 28
10827 Berlin
KI-Innovationswettbewerb@lhlk.de

Status

April 2022

Images

peshkov (title, p. 6), Yucel Yilmaz (p. 12, 17, 19) – stock.adobe.com

EXECUTIVE SUMMARY

For Germany alone, it is expected that services and products based on the use of artificial intelligence (AI) will generate revenues of 488 billion euros in 2025 - this would represent 13 percent of Germany's gross domestic product. In important application sectors, the explainability of decisions made by AI is a prerequisite for acceptance by users, for approval and certification procedures, or for compliance with the transparency obligations required by the GDPR. The explainability of AI products is therefore one of the most important market success factors, at least in the European context.

This study was conducted by the accompanying research for the innovation competition "Artificial Intelligence as a Driver for Economically Relevant Ecosystems" (AI Innovation Competition) on behalf of the Federal Ministry for Economic Affairs and Climate Action. The study is based on the results of an online survey and in-depth interviews with AI experts from industry and science. The study summarizes the current state of the art and the use of Explainable Artificial Intelligence (XAI) and explains it using practical use cases.

The core of AI-based applications - by which we essentially mean machine learning applications here - is always the underlying AI models. These can be divided into two classes: White-box and black-box models. White-box models, such as decision trees based on comprehensible input variables, allow the basic comprehension of their algorithmic relationships. They are thus self-explanatory with respect to their mechanisms of action and the decisions they make. In the case of black-box models such as neural networks, it is usually no longer possible to understand the inner workings of the model due to their interconnectedness and multi-layered structure. However, at least for the explanation of individual decisions (local explainability), additional explanatory tools can be used in order to subsequently increase comprehensibility. Depending on the specific requirements, AI developers can fall back on established explanation tools, e.g. LIME, SHAP, Integrated Gradients, LRP, DeepLift or GradCAM, which, however, require expert knowledge. For mere users of AI, only few good tools exist so far that provide intuitively understandable decision explanations (saliency maps, counterfactual explanations, prototypes or surrogate models).

The participants in the survey conducted as part of this study use popular representatives of white-box models (statistical/probabilistic models, decision trees) and black-box models (neural networks) to roughly the same extent today. In the future, however, according to the survey, a greater use of black-box models is expected, especially neural networks. This means that the importance of explanatory strategies will continue to increase in the future, while they are already an essential component of many AI applications today. The importance of explainability varies greatly depending on the industry. It is considered by far the most important in the healthcare sector, followed by the financial sector, the manufacturing sector, the construction industry and the process industry.

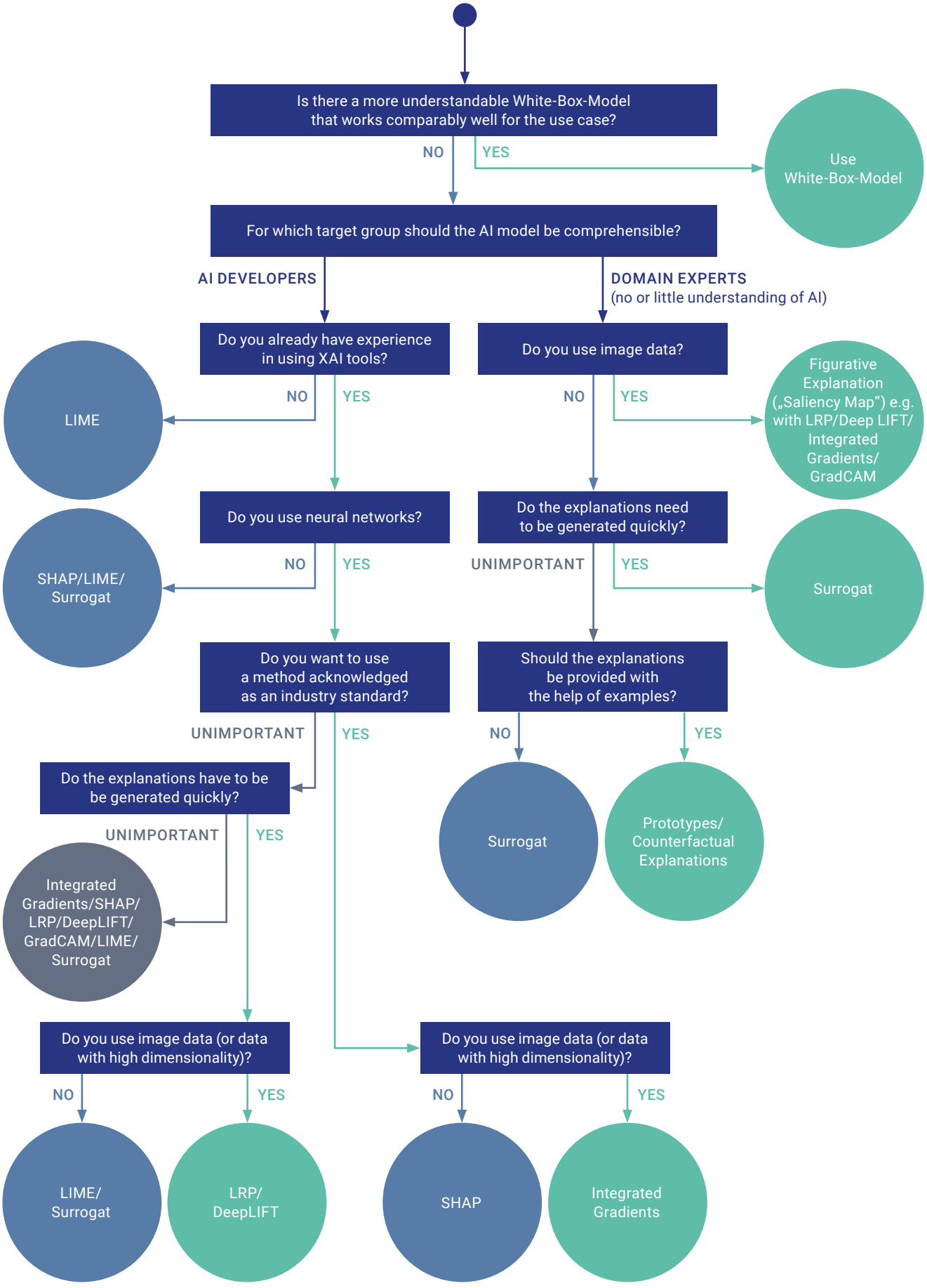
Four use cases were analyzed in more detail through in-depth interviews with proven experts. The use cases comprise image analysis of histological tissue sections as well as text analysis of doctors' letters from the health care domain, machine condition monitoring in manufacturing, and AI-supported process control in the process industry. Among these, model explanations that make the model-internal mechanisms of action comprehensible (global explainability) are only indispensable for the process control case as a strict approval requirement. In the other use cases, local explainability is sufficient as a minimum requirement. Global explainability, however, plays a key role in the acceptance of AI-supported products in the considered use cases related to manufacturing industries.

Furthermore, the use case analyses show that the selection of a suitable explanation strategy depends on the target groups, the data types used and the AI model used. The study analyzes the advantages and disadvantages of the established tools along these criteria and offers a corresponding decision support (see Figure 1). Since white-box models are self-explanatory in terms of model action mechanisms and individual decisions, they should be preferred for all applications that place high demands on comprehensibility - whenever possible. Especially if they perform similarly well, or at least sufficiently well, compared to black-box models.

It can be assumed that with the increasing use of AI in business, the need for reliable and intuitive explanation strategies will also increase significantly in the future. In order to meet this demand, the following technical and non-technical challenges currently need to be overcome:

- New and further development of suitable "hybrid" approaches that combine data- and knowledge-driven approaches, or white- and black-box modelling approaches respectively.
- Consideration of aspects from behavioural and cognitive science - such as the measurability of the quality of an explanation from the user's point of view, automated adaptations of explanations to users, explainability of holistic AI systems - in order to improve explainable AI systems
- Definition of application and risk classes from which the basic necessity of an explanation for given use cases can be derived
- Definition of uniform requirements for the explainability of AI and thus the creation of clear regulatory specifications and approval guidelines corresponding to the application and risk classes
- Creation of approval and (re)certification frameworks for systems continuously learning during operational deployment
- Provision and implementation of comprehensive education and training programs for examiners and inspectors to verify the explainability of AI.

Guidance on the use of the most common strategies and tools for explainable AI („XAI tools").



CONTENTS

EXECUTIVE SUMMARY 3
1 INTRODUCTION 10
2 EXPLAINABILITY OF ARTIFICIAL INTELLIGENCE: OBJECTIVES, CLASSIFICATION AND CONCEPTS 16
2.1 Overarching goals and classification of explainable artificial intelligence 16
2.2 Basic concepts of explainable artificial intelligence 18
2.2.1 Transparency 18
2.2.2 Explainability 20
2.2.3 White-Box and Black-Box Models 21
3 ADVANTAGES AND DISADVANTAGES OF ESTABLISHED STRATEGIES AND TOOLS FOR EXPLAINABLE AI 24
3.1 Integration of prototypes 24
3.2 Integration of external knowledge bases 25
3.3 Surrogate models (substitute models) 25
3.4 Counterfactual Explanations 26
3.5 LIME (Local Interpretable Model-Agnostic Explanations) 26
3.6 SHAP (SHapley Additive exPlanations) 27
3.7 Attribution Methods 28
3.7.1 CAM / Grad-CAM / Grad-CAM++ (Gradient-weighted Class Activation Mapping) 28
3.7.2 LRP (Layer-Wise Relevance Propagation) 28
3.7.3 IG (Integrated Gradients) 29
3.7.4 DeepLIFT (Deep Learning Important FeaTures) 29
3.7.5 Guided Backpropagation und Deconvolution / DeconvNet 29
3.7.6 Activation Maximization 30
3.7.7 Sensitivity analysis 30
4 THE CURRENT USE OF EXPLAINABLE AI IN INDUSTRY AND SCIENCE 34
5 USE CASES FOR EXPLAINABLE AI 42
5.1 Use cases in healthcare 42
5.1.1 Use case: AI-supported image analysis of histological tissue sections 43
5.1.2 Use case: AI-supported text analysis of medical reports 47
5.1.3 Regulation and certification in healthcare 51
5.2 Use cases in manufacturing 52
5.2.1 Use case: AI-supported machine condition monitoring 53
5.2.2 Use case: AI-supported process control in the process industry 57
5.2.3 Regulation and certification in the manufacturing industry 62
5.3 Overall consideration of the use cases 64
6 PRACTICAL FIRST STEPS: ORIENTATION GUIDE FOR SELECTION OF EXPLANATORY STRATEGIES 68
7 CHALLENGES AND NEEDS FOR ACTION FOR THE ESTABLISHMENT OF EXPLAINABLE KI 74
7.1 Technical challenges and need for action 74
7.2 Regulatory challenges and need for action 76
8 CONCLUSION 82
A OVERVIEW OF AI METHODS AND MODELS 88
BIBLIOGRAPHY 94

1 INTRODUCTION

1 INTRODUCTION

Today, artificial intelligence (AI) applications are mostly based on algorithms, processes and models that are multi-layered and intertwined. As a result, the decision-making process of AI-systems is in many cases no longer comprehensible to humans - including the developers of AI.

While in some areas of application, such as product recommendations in the entertainment sector, these circumstances are not perceived as problematic, comprehensibility can be very decisive elsewhere when it comes to using AI products in practice: On the one hand, whenever a certain degree of "explainability" of algorithmic systems is indispensable for regulatory authorities, e.g. in the healthcare industry. On the other hand, if the the target customers do not accept the AI product without a minimum of explainability, for example in automated securities trading.

In Germany and Europe, the extent to which explainability is required for individual approval or certification of AI-supported systems is not yet conclusively clarified in many application sectors - which is a barrier to innovation for companies with these target markets. In view of the forecast sales of AI-based services and products of 488 billion euros for 2025 (eco - Verband der Internetwirtschaft e.V. 2019), this is also of economic relevance. The European Commission, which is pursuing a "risk-based approach" with regard to a future legal framework for AI, takes the view that legal explainability requirements should primarily depend on the criticality of the application (European Commission 2020). However, a concrete definition of what exactly characterises AI systems with high risk potential and what degree of explainability is appropriate is still pending on the part of the EU Commission (Remark: the editorial deadline of the original study predated the proposal of the EU commission for a regulatory framework on AI published on april 21st 2021).

In Germany and Europe, the extent to which explainability is required for individual approval or certification of AI-supported systems is not yet conclusively clarified in many application sectors - which is a barrier to innovation for companies with these target markets.

If possible errors in an AI system are associated with potentially serious or fatal consequences for the life and limb of persons, such as in the health sector, then a "certain" level of explainability must in fact already be ensured today in order to meet the basic requirements for the approval of AI-supported products. However, in this respect, much is left to the discretion of the authorising authorities, as no clear requirements for explainability can be derived from the laws to date. The Medical Devices Ordinance formulates requirements for "risk management", for example, but at the same time does not specify what this means in concrete terms in terms of explainability. In the health sector and in many other application sectors, e.g. in autonomous driving and in the financial economy, there is currently a great need for concretisation which, on the one hand, should be formulated in a technology-neutral way and, on the other hand, must clarify open questions with regard to learning systems.

For less critical AI applications, such as music or film recommendations on entertainment platforms, there are no regulatory requirements regarding explainability. Here, the acceptance of the customers alone is decisive. However, end users also increasingly demand a certain degree of comprehensibility - even if so far less in the consumer sector than in the B2B[1] segment. The need for explainable AI on the company side becomes most apparent when incorrect decisions by AI systems might cause potentially high economic damage (as e.g. in the maintenance planning of expensive machines or systems). In the European or German consumer market, a medium- to long-term increase in demand for explainable AI is also conceivable in principle. Corresponding technical progress in combination with the fact that the General Data Protection Regulation has

1 "Business-to-business" means business relationships between two or more Company

enshrined transparency obligations in law, can change or strengthen citizen awareness in this regard. This study was conducted by the accompanying research for the innovation competition "Artificial Intelligence as a Driver for Economically Relevant Ecosystems" (AI Innovation Competition) on behalf of the Federal Ministry for Economic Affairs and Climate Action. The study is aimed at providers and developers of systems who would like to provide products based on AI and are currently faced with the question of what requirements exist for the explainability of a system and how these can be addressed.

The goals of the study is to classify and to define explainable AI, as well as to provide the advantages and disadvantages of established explanatory strategies. Furthermore, it is the aim to analyse the current use of explainable AI in business and science and to illustrate this based on practical use cases. Finally, we aim to provide an orientation guide for the selection of explanatory strategies and to identify the challenges and needs for action for the realization of explainable AI.

Methodology of the study

The study is based on a survey of 209 representatives from business and science with a connection to the topic of artificial intelligence, a series of interviews with experts, and an extensive literature search. The participants were recruited from the ranks of the members of the German KI-Bundesverband e.V. and from the projects of the BMWK technology programmes KI-Innovationswettbewerb, Smarte Datenwirtschaft, PAiCE and Smart Service Welten.

A total of 209 people took part in the survey, which was implemented as an online multiple-choice questionnaire, from July to October 2020 (72 percent company employees, 26 percent representatives from academia and two percent "others"). Of the company representatives, 70 percent classified themselves as small or medium-sized enterprises (SMEs, with no more than 250 employees) and 30 percent as large companies. 77 percent described themselves as AI providers or developers. Around 23 percent stated that they were AI users or AI users.

Allocation of participants according to target and application industries (multiple answers were possible)*.

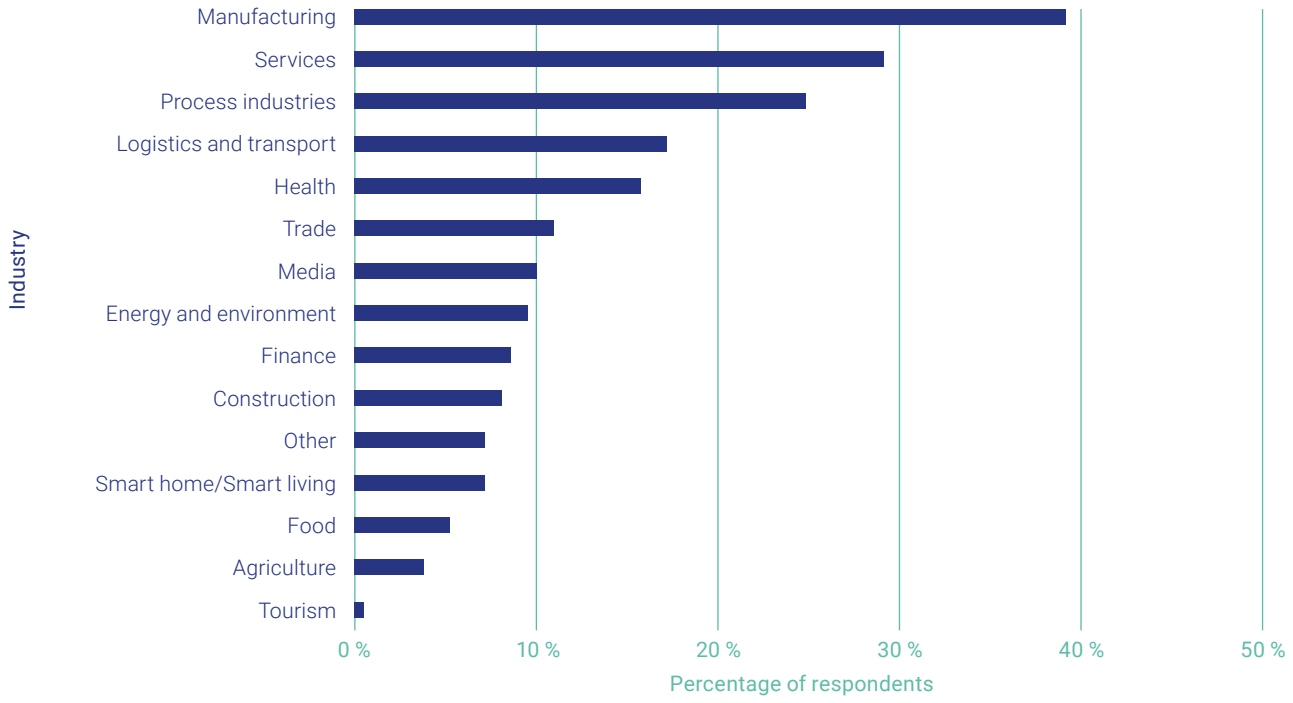
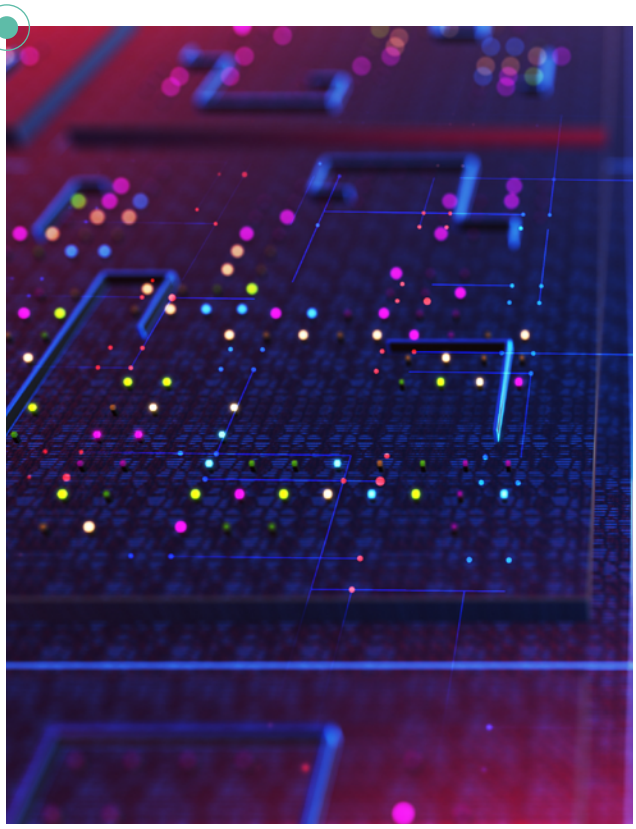


Figure 2 - Allocation of participants by target and application industries; n=209

* Some industries or fields of activity that could not be selected in the questionnaire, e.g. IT/software or public administration, were indicated by several persons as application industries and classified in the figure under „Other".

Several questions in the survey were asked in direct relation to the target or application industries indicated by the participants (see Figure 2). This approach was chosen to make the reference frame for the assessments clear for the respondents. Respectively, only assessments by people with domain insight are taken into account for the resulting comparisons among industries. For industries that only a small minority of participants indicated as their field of application, the corresponding results must be considered less reliable due to the greater influence of individual opinions.

The interviews with the experts were conducted from November 2020 to February 2021. The interview partners all have a professional connection to artificial intelligence and cover the areas of research, business, standardization and licensing. In individual conversations via web conference, guideline-based open interviews were conducted with the experts on the topics “technical implementation of explainable AI”, “established explanatory tools” and “technical and regulatory challenges”.



Overview of the study

The study is structured as follows:

- The most important concepts from the relevant literature are compiled in order to facilitate access to the professional discourse. Chapter 2 defines the central concepts such as transparency, white-box and black-box models as well as decision and model explanations and, respectively, local and global explainability.

- The established explanatory strategies and tools that represent the state of the art are presented in Chapter 3 and discussed in terms of their potential applications and practical benefit.

- Important fields of usage in terms of application industries, model and process categories, target groups as well as data types and implementation possibilities were identified in the course of the survey conducted. The results are presented and discussed in chapter 4.

- Based on four use cases from the healthcare sector (image analysis of histological tissue sections, text analysis of medical reports), manufacturing (machine condition monitoring) and process industry (process control), overarching goals and concrete explainability requirements from the perspective of relevant target groups are identified, compared with each other and corresponding solutions are described. The use cases are described in chapter 5.

- In the course of the expert interviews, advantages and disadvantages as well as fields of application of established explanatory strategies and tools were discussed. From this, a compact orientation guide was generated in the form of a decision tree, which can be found in Chapter 6.

- Essential technical and regulatory challenges and needs for action for the realization of explainable AI systems were identified in the context of guideline-based interviews with experts and discussed in Chapter 7.

The team of authors would like to thank the experts who made themselves available for the interviews. At the same time, we would like to thank all participants of the survey and the Bundesverband KI e.V., and in particular its managing director Daniel Abbou, for the cooperation in approaching the members. The responsibility for all statements made in this study lies exclusively with the team of authors.

The authors would like to thank the experts for their participation in the interviews:

- Dr. Tarek Besold, DEKRA Digital GmbH
- Dr. Richard Büssow, Industrial Analytics IA GmbH
- Christian Geißler, Technical University Berlin
- Prof. Dr. Martin Hirsch, University of Marburg, Ada Health GmbH
- Prof. Dr. Marco Huber, University of Stuttgart, Fraunhofer IPA
- Prof. Dr. Alexander Löser, Data Science Research Center, Beuth University of Applied Sciences Berlin
- Prof. Dr. Axel-Cyrille Ngonga Ngomo, University of Paderborn
- Dr. Christoph Peylo, Bosch Center for Artificial Intelligence, Robert Bosch GmbH
- Prof. Dr. Philipp Rostalski, University of Lübeck
- Prof. Dr. Ute Schmid, University of Bamberg, Fraunhofer IIS
- Gerald Spyra, Ratajczak & Partner mbB
- Thomas Staufenbiel, Gestalt Robotics GmbH
- Martin Tettke, Berlin Cert GmbH
- Prof. Dr. Leon Urbas, Dresden University of Technology
- Betty van Aken, Data Science Research Center, Beuth University of Applied Sciences Berlin

We would also like to thank the Bosch Center for Artificial Intelligence for the information provided. Last but not least, we would like to thank the Fraunhofer IPA, in particular Nina Schaaf, for the discussions on the topic of explainable artificial intelligence and for the cooperation with regard to the coordination and the differentiation of this study and the one carried out at Fraunhofer IPA on explainable artificial intelligence.

2 EXPLAINABILITY OF ARTIFICIAL INTELLIGENCE: OBJECTIVES, CLASSIFICATION AND CONCEPTS

2 EXPLAINABILITY OF ARTIFICIAL INTELLIGENCE: OBJECTIVES, CLASSIFICATION AND CONCEPTS

2.1 Overarching goals and classification of explainable artificial intelligence

The motivation for the development and implementation of explainable AI differs depending on the use case and the interests of the target groups of explainable AI. Nevertheless, overarching goals of explainable AI can be formulated - based on (Arrieta et al. 2019) - and pursued either individually or in combination:

1. **Check plausibility of causal relationships:** With explainable AI, patterns discovered by AI should be additionally tested for their validity and plausibility. A frequent motivation of users is "finding" or mapping causal relationships.

2. **Test transferability:** AI models are usually trained to solve a specific task, e.g. "find all pictures with a cat". Explainability should help to estimate the transferability of the found solution to new tasks (e.g. "find all dogs"). This helps to determine the scope and limits of an AI.

3. **Increase information gain:** In order to be able to use an AI system at all (e.g., as a decision support system), information about the basics of the decision-making process should be provided in an understandable and simple, but not oversimplified, form.

4. **Determine confidence:** The AI system should be checked for robustness (preservation of the system's goodness-of-fit criteria when assumptions are not met or statistical outliers play a role), stability (similar data yield similar results), or reproducibility (same result when run multiple times) to identify vulnerabilities and areas of validity.

5. **Fairness testing:** With this goal in mind, explainable AI will be used to test a model for fairness, specifically to detect any biases (i.e., systematic errors) that may exist in the database.

6. **Improve interaction possibilities:** Explainable AI should support users - especially those with little AI expertise - to interact directly with the AI system, for example to improve its decision-making or the comprehensibility of explanations, e.g. by providing alternative explanations (summary of "interactivity" and "accessibility" from (Arrieta et al. 2019)).

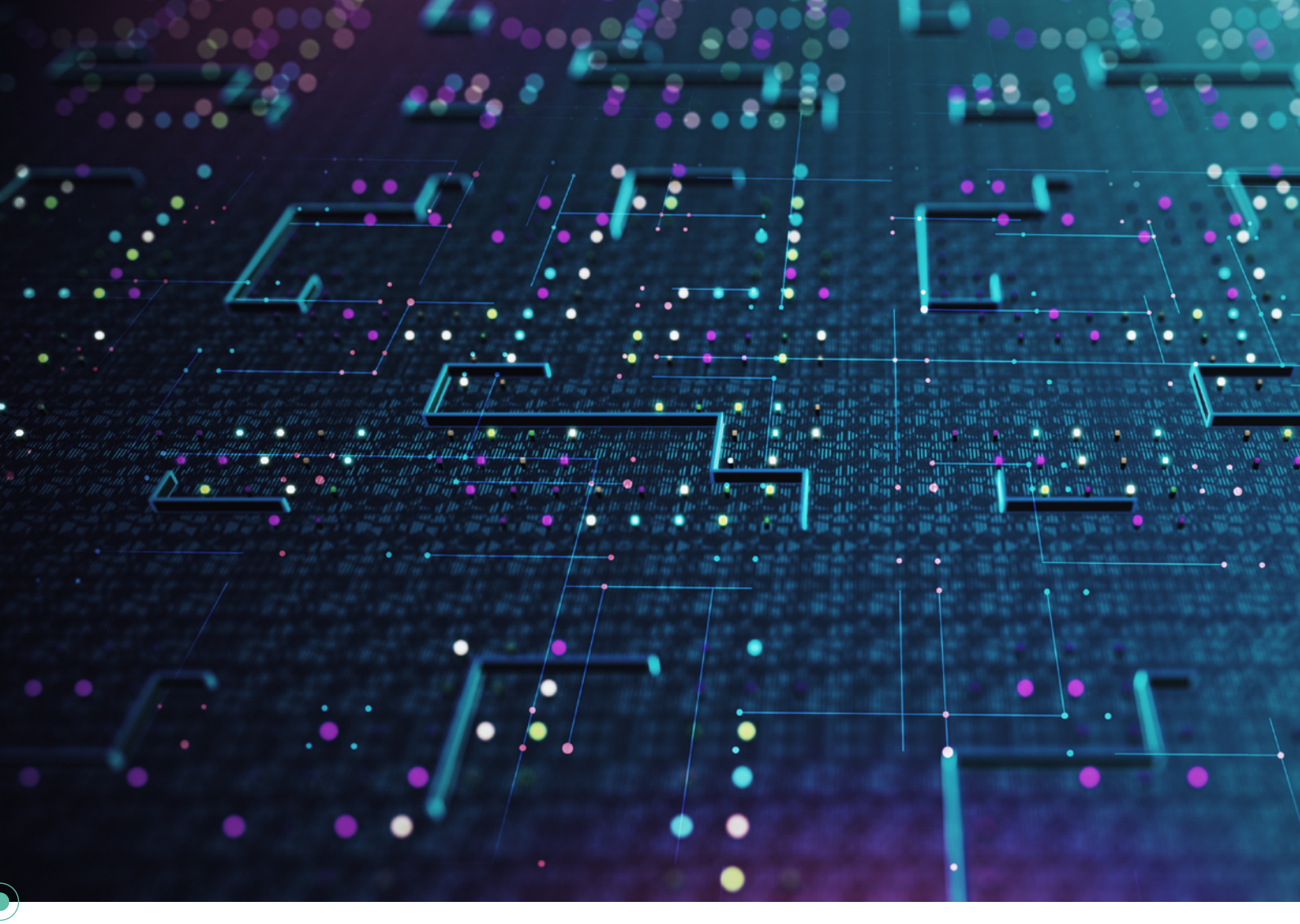
7. **Increase privacy awareness:** Potentially, explainable AI can provide users with insight into the data collected and stored, leading to an increased awareness of privacy aspects.

8. **Clarify responsibilities:** Explainable AI can be used to clarify responsibilities and liability issues. For example, it could be established via a court-appointed expert that large amounts of biased data have been deliberately introduced into a system in order to influence it.

An additional goal that is also frequently formulated as a motivation in the literature (Ribeiro et al. 2016; Arrieta et al. 2019) is the establishment of trust or trustworthiness, which, however, is no longer seen only as an overarching goal of explainability, but as a general development goal of AI systems.

According to the European Union's High-Level Expert Group on AI guidelines for trustworthy AI (High-Level Expert Group on AI 2019), explainability per se helps to establish trust. The seven pillars that should support trust in AI are:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency (e.g. traceability, explainability and communication)
5. Diversity, non-discrimination and fairness
6. Social and environmental well-being
7. Accountability



In the context of the seven pillars, "transparency" refers to the property of an AI to have a traceable and, above all, explainable decision-making process, which is communicated to the user via appropriate information.

At the same time, the mechanisms of action within the seven pillars are not entirely independent of each other. If explainability (as one of the core aspects of "transparency") is given, then this also has a direct effect on the points 1, 2 and 5 from above:

- For example, if AI systems are vulnerable to data bias[1] in real-world data, transparency and explainability in (semi-) autonomous systems can greatly facilitate human oversight. Also, in decision support systems, the risk of decisions being made without adequate human review can at least be reduced if a certain degree of explainability is inherent.
- From a developer's perspective, the vulnerability of systems to data bias can be better addressed if transparency is ensured, which will also improve the technical robustness and saftety of the systems.
- Transparency is also a prerequisite to enable equal treatment of individuals and to identify potentially discriminatory decisions of algorithmic systems (non-discrimination and fairness[2]).

Explainability is also an important aspect that can contribute to the acceptance of AI. Basically, the acceptance of a technology is determined by its voluntary, active and targeted use.

To date, there is no generally accepted, and consequently no uniform, taxonomy for explainable artificial intelligence, which is why the following section will go into more detail about which concepts and terminology are used as the basis for the study.

1 Here and in the following, the term "data bias" or "bias" refers to to an application-neutral and statistics-related understanding of the expression, i.e. a general systematic deviation is meant.

2 Nevertheless, testing fairness is almost always a challenge, especially because of the difficulty of selecting appropriate metrics.

2.2 Basic concepts of explainable artificial intelligence

Despite sometimes conflicting labels, there is widespread consensus (Lipton 2016; Gilpin et al. 2018; Arrieta et al. 2019) on the basic distinction between two concepts, namely.

- Transparency[3]
- and explainability (mostly in the form of post hoc explainability).

The terminology in this study is based on (Arrieta et al. 2019), and the presentation is inspired by (Lipton 2016).

If an AI model is transparent - whereby transparency is to be understood explicitly as a property here - it can also be referred to as a "white-box" model, provided that the input data are comprehensible. With such models, in particular the algorithmic mechanisms for generating the model are comprehensible. A detailed definition of transparency follows in section 2.2.1.

Explainability, on the other hand, is about actively providing a target person with an understandable rationale that allows them to comprehend the outcome of an AI model.

The perception and the the level of knowledge of the target person, but also the orientation of the question, must necessarily be taken into account when creating explanations. A detailed definition follows in section 2.2.2.

> If an AI model is transparent, it can also be referred to as a "white-box" model, provided that the input data are comprehensible. Explainability, on the other hand, is about actively providing a target person with an understandable rationale that allows them to comprehend the outcome of an AI model.

2.2.1 Transparency

Transparency is treated as a model property in the following. If the transparency of a model is given, it is self-explanatory under the assumption of comprehensible input variables[4]. The property of transparency can be further subdivided into the three different manifestations of "simulability", "decomposability" and "algorithmic transparency" (Lipton 2016). Here, hierarchical dependence is often assumed in the literature (Arrieta et al. 2019), such that the simulability of a system implies its decomposability and its algorithmic transparency. Correspondingly, the decomposability of a system also establishes its algorithmic transparency. Consequently, assuming explainable input data, a model is already considered transparent if it only satisfies the property of algorithmic transparency. A model reaches the highest level of transparency if it satisfies the property of simulability and thus also fulfills the two other properties.

A system is simulatable if even a person can or could reproduce the decisions of the underlying algorithm in a reasonable amount of time by manually performing the individual steps required to bring about a decision.

Example: When manually traversing different paths of a not too large decision tree based on comprehensible input variables, a person can check for himself in each node, whether an individual property of input data or an attribute is satisfied or not. If there are no more attributes to check, the person has reached a "leaf" of the decision tree, which represents the result.

3 Here, the term transparency refers to a different concept than what was previously referenced in the Guidelines for Trusted Artificial Intelligence (High-Level Expert Group on AI 2019).

4 In this context, the literature also frequently refers to "interpretable" models. However, this term has been avoided here because "interpretability" is often used contradictorily in the relevant literature.

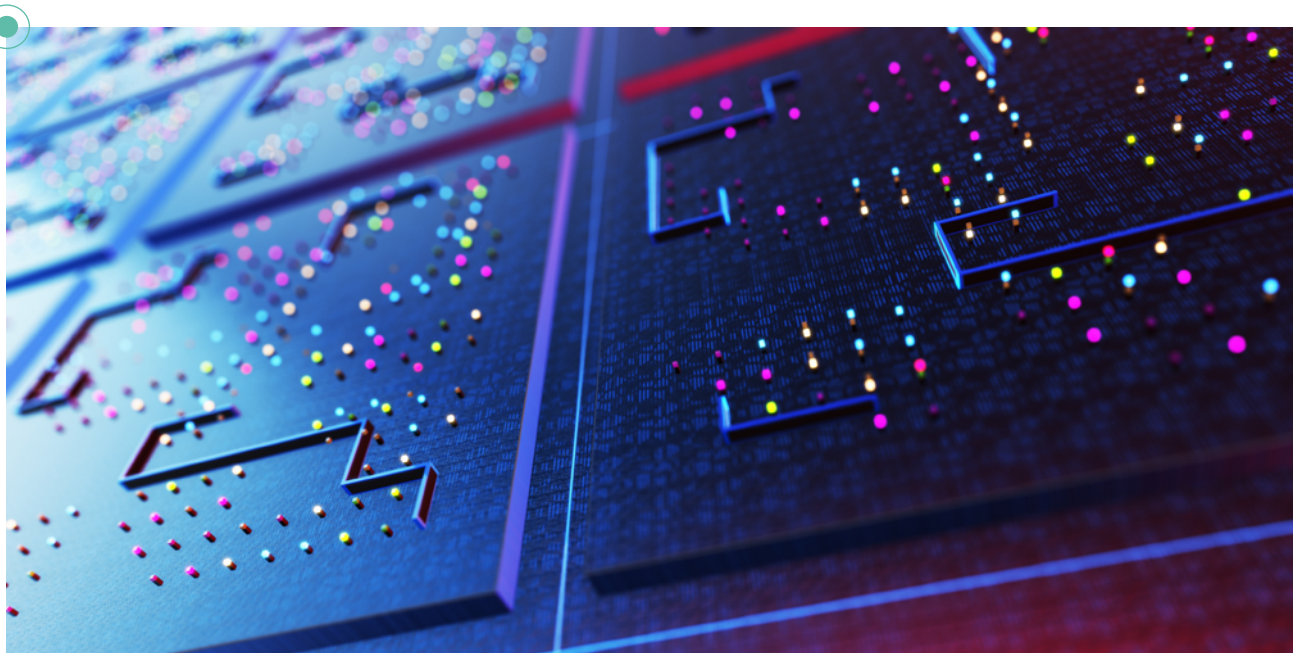
In a decomposable system, the individual components (input data, parameters, model levels, calculations, etc.) can or could be provided with an intuitive description so that their functions in the overall system can be closely understood.

Example: In decision trees, it is specified for each node which attribute is currently being tested and which characteristic is required in each case for the selection of the following subpaths. Individual model levels can be characterized by the description of the nodes they contain. The intuitive description and comprehensibility of the input variables must be ensured during system design in order to be able to describe a decision tree as decomposable, i.e. the input variables themselves must not be opaque constructs consisting of many variables.

Algorithmic transparency refers to the actual learning process or the generation of models. What matters here is whether it is possible to understand how a model is generated in detail and how, during the training phase, possible situations are dealt with that the algorithm in question might be confronted with (in terms of unknown input or training data). In this context, algorithmic transparency is really only about the properties of the algorithm and not about concrete model features or data.

Example: In the case of linear regression, in which a linear model is fitted to a point cloud consisting of measured values, e.g. by means of the method of least squares (standard mathematical procedure for regression analysis), it is possible to understand in detail how the unique result is determined. With certain assumptions about statistical distributions, additional statistical statements can be made about the determined result. In any case, the resulting linear model is always unambiguous, the convergence is reliable and the limits are well known (e.g. susceptibility of the classical least squares method to statistical outliers).

The example of the decision tree and the linear regression model shows that these AI models fulfill basic transparency properties. For other model types, this is not the case even for small-dimension model instances still usable for practical applications. For example, in order to consider neural networks in image processing decomposable, the function of each individual node and each individual layer in the network would have to be clearly describable and referable to the result. In such a system, one node might be “responsible” for detecting horizontal lines in the image, another might detect vertical lines, etc. This description would have to be available for the whole network, so that deeper layers based on the results of the previous ones could also be explained. Therefore, the property of decomposability does not seem to be satisfiable for neural networks. Accordingly, the interconnectedness and multilayeredness that usually characterize neural networks also have the effect that model variants of neural networks - according to the above definition - are not considered simulatable. In the literature, the lack of algorithmic transparency of neural networks is often attributed to the fact that



common training methods for unknown input or training data usually do not lead to unique solutions (Lipton 2016). This is because the error or loss surfaces of the loss function are usually difficult to analyze given the problem structure (Arrieta et al. 2019; Kawaguchi 2016; Datta et al. 2016) and solutions can only be approximated using heuristic optimization methods (Arrieta et al. 2019). According to (Arrieta et al. 2019), in general, the accessibility of a model to appropriate mathematical analyses and methods is the key criterion for the algorithmic transparency of a model.

A discussion of the transparency properties of other models in terms of algorithmic transparency, simulability, and decomposability can be found in (Arrieta et al. 2019).

2.2.2 Explainability

Since the aforementioned transparency is not achievable for various models such as neural networks, which are consequently not self-explanatory, the alternative concept of "explainability" is applied to them when necessary. In this context, it is normally specified whether the explainability concerns a particular decision or a model as a whole. On the other hand, no general rule can be derived as to how a concrete explanation should be designed and how much knowledge it conveys to the target person. In the example of image processing with neural networks, one already speaks of an explanation of a decision when certain areas in the input image, which were most significant for the classification of an image, are highlighted in color for the target person. In this case, not every single step of the algorithm is explained, but the target person is pointed to the data that was most significant for the individual decision. Alternatively, an explanation can be represented by a textual description, e.g. "This picture shows a dog, because four legs, a snout, fur and a tail were recognized."

Overview of white-box/black-box nature of models used for machine learning.

AI Model	Transparency			White box/ Black box	Post-hoc analysis necessary?
	Simulatability	Decomposability	Algorithmic transparency		
Neural networks	X	X	X	Black box	Necessary: Tools in chapter 3
Ensemble models (e.g. Tree Ensembles)	X	X	X	Black box	Necessary: Tools in chapter 3
Support vector machines	X	X	X	Black box	Necessary: Tools in chapter 3
Bayesian networks	(√)	(√)	√	White box*	Not necessary
Linear/logistic regression Models	(√)	(√)	√	White box*	Not necessary
Decision trees	(√)	(√)	√	White box*	Not necessary

Table 1: Overview of white-box/black-box nature of models used for machine learning (based on (Arrieta et al. 2019)).

* Applies in the case of comprehensible input parameters and generally in the case of decomposability.

Basically, there are two types of explanations:

- Explanations of individual decisions or decision explanations that help to concretely trace individual, data-related decisions (so-called local explainability or data explainability).
- Explanations of models or model explanations that help to understand the interdependencies of AI models (so-called global explainability or model explainability), e.g. linear or general functional relationships between input and output variables.

It is a "post hoc" explanation when an appropriate analytical tool is applied "retrospectively" to generate an explanation - i.e. after an individual decision making (in the case of decision explanations) or after model training (in the case of model explanations). Post hoc explanations can theoretically be generated regardless of whether the model is transparent or 'opaque' - at least if the analysis tool used is suitably flexible. Normally, however, such explanations are only required in order to establish a certain degree of comprehensibility for opaque models ("black box").

2.2.3 White-Box and Black-Box Models

Based on the presented paradigm of transparency, models can be assigned to the class of black-box models if none of the three properties - simulatability, decomposability or algorithmic transparency - are fulfilled. Conversely, models that satisfy at least the lowest of those three transparency levels (algorithmic transparency) and use comprehensible input variables will be referred to in the following as white-box models.

An overview of whether the currently frequently used AI models fulfill the properties of simulatability, decomposability and algorithmic transparency, and thus can be considered white or black box accordingly under the assumption of comprehensible input variables, can be found in Table 1. The classification of the individual models essentially follows the concept of (Arrieta et al. 2019).

It can be seen that the division into white-box and black-box models succeeds very well with the help of the transparency levels. Although each nominal white-box model can potentially lose the two properties of simulability and decomposability e.g. when its dimension is too high and thus the meaning of certain model layers or variables cannot be assigned intuitively anymore, the algorithmic transparency is maintained in any case. This distinguishes white-box models crucially from their black-box counterparts, which do not satisfy any of the three properties even for small-dimension models that are still of actual practical use in applications. Bayesian networks are an interesting special case. This class of models has the advantage that statistical information about training data (e.g. "density" of training data in the data space) can be taken into account when computing confidence values. This means that Bayesian networks not only provide the decision itself, but also provide quantitative statements, e.g. about how likely the occurrence of an event is. This property is also maintained if the requirements for simulability and decomposability are not met.

In contrast to self-explanatory white-box models, black-box models - for example, neural networks - require the use of an additional strategy to make the model comprehensible or to explain it. This is a "post hoc" analysis, when an appropriate explanatory tool is applied to the AI model in retrospect, i.e., after the decision has been made or the AI model has been trained. In the following chapter, various explanatory strategies are presented, most of which can be referred to as post hoc analysis tools. Other approaches discussed add certain components to the AI models themselves that allow explanations to be extracted from the extended models.

3 ADVANTAGES AND DISADVANTAGES OF ESTABLISHED STRATEGIES AND TOOLS FOR EXPLAINABLE AI

3 ADVANTAGES AND DISADVANTAGES OF ESTABLISHED STRATEGIES AND TOOLS FOR EXPLAINABLE AI

With regard to explanatory strategies, a distinction can be made between approaches that provide "model explanations" and approaches that provide "decision explanations". A model explanation provides information about the concrete functioning of the model. A decision explanation provides reasons that led to a single decision of the AI model. An AI model can be self-explanatory per se (white box) or - often due to appropriate extensions of black-box models - generate explanations simultaneously with the decision. Alternatively, explanations can be provided after the decision (post hoc) by an additional (post hoc) analysis tool. The latter approach specifically addresses black-box models and the improvement of their comprehensibility.

White-box models - for example linear and logistic regression models, decision trees or Bayes nets - are self-explanatory in terms of model-action mechanisms (due to their directly comprehensible functioning) and with regard to their decisions. Consequently, the white-box model can be used for the concrete task (e.g. classification, regression or clustering) as well as for providing explanations.

The explanatory strategies presented below are only a limited selection; many more methods are used and discussed in research and practice. The selection includes the ten

explanatory tools whose associated first scientific publication has at least 500 citations according to Google Scholar (as of December 2020), as well as established methods that were additionally named by the experts during the interviews. For each explanatory strategy presented here, a brief discussion is provided that includes an easy-to-understand example, a brief description of the technical background, important advantages and disadvantages, and references to further reading.

With regard to explanatory strategies, a distinction can be made between approaches that provide "model explanations" and approaches that provide "decision explanations". A model explanation provides information about the concrete functioning of the model. A decision explanation provides reasons that lead to a single decision of the AI model.

3.1 Integration of prototypes

Explanation Type:
Decision explanations; a model provides both a decision and an explanation

Applicable to:
all models, independent of their concrete implementation (but with focus on classification problems); image and text data as well as numerical data

Example:
An AI model is designed to assign patients to a clinical picture on the basis of their symptoms. For each clinical picture, e.g. cold, flu or pneumonia, a prototype is created.

This prototype functions as a kind of fact sheet that summarizes the most common symptoms. The prototypes can be created on the basis of the symptoms of many different patients suffering from the corresponding disease. For each patient to be classified, a profile is also created containing the symptoms, e.g. cough, fever and cold. This is then compared with the representations of the individual classes (clinical pictures) and the most similar one is selected.

Technical background:
Using prototypes is about creating representations of individual classes. These representations can be, for example, data points from the training base that describe the respective class well or artificially generated representations that include the characteristic features for the respective class. These artificial representations can be generated using generative networks such as Generative Adversarial Networks or Variational Auto-encoders. It is also possible that a class is characterized by several prototypes. To provide an explanation, the most similar prototype must be found for a classification result. For

this purpose, for example, a K-Nearest-Neighbor search[5] can be used. Finally, the user can compare the prototype used for the concrete class as well as the data values entered and thus understand on which basis, i.e. according to which similarities, the decision was made by the AI model.

Pros:
- Number of prototypes can be freely selected
- Intuitive and easy to understand
- Independent of AI model and data types

Cons:
- Number of prototypes needed may be unclear
- For artificially created prototypes: may not be realistic

(Molnar 2019; Barbalau et al. 2020; Li et al. 2017)

3.2 Integration of external knowledge bases

Explanation Type:
Decision explanations; a model provides both the decision and the explanation

Applicable to:
All models, independent of their concrete implementation (focus on classification problems); text data only (knowledge base required)

Example:
The PubMed database contains numerous medical articles that describe, among other things, specific diseases and their symptoms. By using this knowledge base, an AI model can learn the relationships between symptoms and diseases. If the model is then confronted with symptoms from patients, such as a severe swelling and a blood clot, the model can then use this knowledge to learn the correlations between symptoms and diseases. If the patient has ankle effusion and pain on exertion, the result of the model could be "suspected ligament rupture". At the same time, a reference is made to one or more articles from PubMed in which precisely these symptoms and the derivation of the corresponding clinical picture are described and clearly highlighted for the user.

Technical Background:
In this approach, the AI model (e.g. neural networks) is combined with external knowledge bases. The knowledge base is already used during the training of the AI model to create the model. Knowledge bases can be, for example, online publications on a certain topic, reference books or Wiki-pedia. The goal here is to learn correlations from the knowledge bases and to be able to justify decisions made by the AI model with concrete entries in the knowledge base.

Pros:
- Easily comprehensible: Reliability of the publications on which the decisions are based can be easily checked
- Combination of several knowledge bases possible

Cons:
- Depending on the quality (and existence) of the knowledge base
- Independent development of a qualitative knowledge base is very time-consuming

(van Aken et al. 2021; Holzinger et al. 2017)

3.3 Surrogate models (substitute models)

Explanation Type:
Related to the original model, neither model nor decision explanations, as a new model is created; post hoc

Applicable to:
All models, regardless of their concrete implementation; image and text data as well as numerical data

Example:
An AI model that is not easy to understand is trained, e.g. a Support Vector Machine, to forecast the daily rental figures for winter sports equipment. Many factors flow into the model, such as the time of year, the weather report, the times of the school holidays and the day of the week. Now a simpler model, such as a decision tree, is trained, whose decisions are more comprehensible, but which cannot represent the complete complexity of the original model. Thus, the forecasts of the second model are often less accurate, but generally valid regularities can be derived, such as: "Fewer skis are rented

5 K-Nearest-Neighbor is a method to assign similar further data points to a given data point. The similarity can be determined in different ways.

when it is foggy" or "The number of rentals is significantly lower on Mondays than on Sundays".

Technical Background:
Surrogate model building is about building a second model, such as a linear model or decision tree, that is more tractable than the original black box model and can thus be used to explain the decision-making. Building on the inputs and outputs of the original model, the prediction function of the surrogate model is derived. Thus, rules can be extracted from trained neural networks and comprehensible decision trees can be created based on them, for example with the algorithm TREPAN (Touretzky 1996). Especially for image data, the creation of such trees is not trivial.

Pros:
- Very flexible: original and surrogate model can be freely selected

Cons:
- Only approximation: representativeness of the surrogate model difficult to measure
- Surrogate model itself can become very complex and less comprehensible
- Not easily applicable to image data

(Adadi und Berrada 2018; Danilevsky et al. 2020; Molnar 2019)

3.4 Counterfactual Explanations

Explanation Type:
Decision explanations; Post hoc

Applicable to:
All models, regardless of their concrete implementation; image and text data as well as numerical data

Example:
When reviewing applications for a rental apartment, the interested parties are evaluated with regard to several criteria and the most suitable prospective tenant is selected. For example, the three factors income, pet ownership and credit rating are considered here. If the AI system now rejects a prospective tenant with an annual income of 40,000 euros, no pets and a positive credit report, a possible explanation would be: With an income of 45,000 euros, the applicant is eligible for a rental. In the same way, the concept can be used in case of a positive decision. Exemplary explanations would be: "If the applicant had a cat, he/she would not be considered as a potential tenant" or "If the credit rating was negative, he/she would not be able to rent the apartment".

Technical Background:
Counterfactual Explanations is a concept that aims to identify the smallest possible change in the input values for a concrete classification result that would lead to a classification in a different class. However, there may be multiple ways to vary the input values such that the classification algorithm arrives at a different result. The following four properties characterize good counterfactual explanations:

- The original classification result and the new one caused by the change of input values are very similar
- As few features as possible should be changed
- Several different explanations can be helpful
- The changed features should be realistic

Practical implementations do not necessarily address all of these characteristics, so it must be checked which are most desirable for one's own use case. For example, the implementation by Wachter et al. focuses only on the first two properties (Wachter et al. 2017). For this purpose, a corresponding loss function is set up, which is optimized with respect to one or more objectives.

Pros:
- Well understandable
- No data or access to the inner model structure required

Cons:
- Proposed change may not be realistic or even impossible in practice
- Multiple, contradictory explanations possible

(Wachter et al. 2017; Stepin et al. 2021; Molnar 2019)

3.5 LIME (Local Interpretable Model-Agnostic Explanations)

Explanation Type:
Decision explanations; Post hoc

Applicable to:
All models, regardless of their concrete implementation; image and text data as well as numerical data

Example:
As a concrete example, the probability of death of a cancer patient is calculated depending on the age of the patient. For a 25-year-old patient, a probability of 45 percent is assumed as the result of the AI model. Now the probability of death is calculated for patients with a similar age, for example 24 (44 percent) and 26 years (46 percent). Using these three values - in practice more are usually used - it is possible to estimate the behaviour of the model within a limited range: e.g. a slight (linear) increase in the probability of death with increasing age. For patients aged 74, 75 and 76, the model may behave differently - for example, it may show a much greater increase in the probability of death with age. In this way, LIME can be used to simplify individual "sections" of what is actually a highly interwoven and complex model, making it easier for the user to understand (Nguyen 2020).

Technical Background:
The basic idea of LIME is to learn a locally approximated, interpretable model for a concrete classification or regression result. This allows a concrete result to be reproduced using a simpler, often linear model, even though the original model is difficult to reproduce. LIME "samples" several results (or decisions) and weights them according to their proximity to the result to be explained. On this basis, a local model can be developed that works well with the samples considered and is comprehensible.

Pros:
- Intuitive and generally easy to interpret
- Quick and easy integration into existing implementations (appropriate framework available)

Cons:
- Problematic for distinctly nonlinear models
- Possibly high computing time for multidimensional data, e.g. image data
- Hardly reproducible due to data sampling (a classification performed several times could be explained differently)

(Nguyen 2020; Ribeiro et al. 2016)

3.6 SHAP (SHapley Additive exPlanations)

Explanation Type:
Decision explanations; Post hoc

Applicable to:
All models, optimizations exist for individual models (e.g. TreeSHAP for random forests); image and text data as well as numerical data

Example:
As an example, the prediction of income based on the three factors age, gender and occupation will be considered. The influence of each factor on a concrete result of the AI system is determined. In order to find out how important age is for the income forecast, a "normal" forecast is first calculated taking age, gender and occupation into account. Then, a forecast is made again, but using only the two factors of gender and occupation. In this way, the difference between the two results - once taking age into account, once not taking age into account - can be calculated afterwards and the influence of the factor "age" can be determined. This process is repeated for the other two factors (Mazzanti 2020).

Technical Background:
SHAP is an approach from game theory. When applying the method, each feature or input value is weighted with respect to a concrete classification result. These weights are also referred to as Shapley Values. The idea behind this is that all possible combinations of features are considered to determine the importance of an individual feature. Each input feature is thus assigned a positive or negative value indicating the influence of the individual feature on the result. The method can be used to generate explanations of decisions. TreeSHAP is a variant of SHAP that can be applied particularly efficiently to tree-based models.

Pros:
- Model-agnostic (tailored SHAP implementations provide high efficiency)
- Very precise explanations possible
- Considered an industrial standard

Cons:
- Explanations not always intuitive
- Possibly high computing times (especially for models with a high number of parameters)

(Molnar 2019; Lundberg und Lee 2017; Mangalathu et al. 2020; Mazzanti 2020; Bhatt et al. 2019)

3.7 Attribution Methods

With the help of so-called attribution methods, the negative or positive influence of parts or areas of the input of an AI model on its output is considered (Sundararajan et al. 2017). The following concrete methods can be assigned to this group: Sensitivity Analysis, LRP, DeepLIFT, Integrated Gradients, Grad-CAM, Guided Backpropagation and Deconvolution. A common example is used to explain how they work. The differences are described in the following technical details.

Example:
Image classification is about identifying image areas that are crucial for the classification result. As an example, objects on an image are to be recognized with the help of a neural network. The two possible classes are "cat" or "dog". After the AI model has delivered a decision, e.g. "cat", the influence of individual pixels and image areas on the concrete decision is examined. For this purpose, the individual components of the neural network - units and layers - are considered in order to "map" the outputs to the input image. The result is a so-called saliency map, in which the pixels and image areas that had a particularly large influence on the animal being recognized as a cat are highlighted.

3.7.1 CAM / Grad-CAM / Grad-CAM++ (Gradient-weighted Class Activation Mapping)

Explanation Type:
Decision explanations; Post hoc

Applicable to:
Neural networks, especially Convolutional Neural Networks (CAM may require the addition of special layers); image data

Technical Background:
CAM is a method for visualizing crucial regions for a concrete classification result of a neural network, in particular Convolutional Neural Network (CNN). The result is a saliency map that can be overlaid on the original image to highlight the regions of interest. To create the saliency map, only the last layers of the network are considered at a time. CAM is not directly applicable to every network architecture; it may need to be adjusted by adding more layers beforehand and then re-training the network.

Grad-CAM is a generalization of the CAM method, does not require re-training of the model, and is applicable to more network architectures. However, one drawback of Grad-CAM is that it cannot detect multiple occurrences of an object in an image. Grad-CAM++ solves this problem, allowing the detection of multiple object instances in one image.

Pros:
- Visualizations correlate with human attention → easily understandable
- Good results in tasks where image objects have to be localizedGute Ergebnisse bei Aufgaben, in denen Bildobjekte lokalisiert werden müssen

Cons:
- Visualizations often too rough for small image objects → only rough validation (quality strongly depends on concrete application)
- CAM: additional layers must be trained

(Zhou et al. 2015; Selvaraju et al. 2019; Chattopadhyay et al. 2017)

3.7.2 LRP (Layer-Wise Relevance Propagation)

Explanation Type:
Decision explanations; Post hoc

Applicable to:
Neural networks; focus on image data

Technical Background:
LRP considers the influence of individual inputs on the result of a classification. The focus here is on non-linear classifiers such as neural networks. Considering image classification, the goal is to find out for individual images which pixels positively or negatively influence the classification result and to what extent. Each input value (here: pixel) is assigned a relevance value. The 'relevance' value indicates how much influence an input value or a unit of the network has on the classification result. The relevance value of the output is the sum of the relevance values of the input values. The output value of the network is thus 'decomposed' into the respective contributions (or the influence) of the input values. The calculation of the relevance of the input values is performed iteratively from the back (last layer) to the front (input layer).

Pros:
- Good quality of explanations even for multilayer models with a high number of parameters
- Declarations can be generated very quickly (in relation to the runtime)

Cons:
- Numerical problems with decomposition possible → possibly misleading visualizations

(Bach et al.; Samek et al. 2019; Shiebler 2017)

3.7.3 IG (Integrated Gradients)

Explanation Type:
Decision explanations; Post hoc

Applicable to:
Neural networks; image and text data and numerical data

Technical Background:
This method is also intended to improve the explainability of neural networks through visualization. One advantage is that the structure of the network does not need to be changed, as may be the case with CAM. In the exemplary case of image data, when IG is used, an image is chosen as the baseline, such as a completely black image. A series of interpolated images are then created 'between' the baseline and the original input, each with little difference between them. On this basis, individual gradients are calculated, which in turn are used to identify interesting areas - i.e. decisive for the classification - in the input image.

Pros:
- Scales well for image processing
- Positive and negative influence of individual input values can be displayed separately
- Usage of a baseline: intuitive approach
- Considered an industrial standard

Cons:
- Correct choice of baseline unclear → widely varying results
- Explanations not always intuitive

(Sundararajan et al. 2017; Bhatt et al. 2019; Google 2020)

3.7.4 DeepLIFT (Deep Learning Important FeaTures)

Explanation Type:
Decision explanations; Post hoc

Applicable to:
Neural networks; focus on image data

Technical Background:
DeepLIFT is an explanatory tool that is used to improve the comprehensibility of neural networks. The method assigns a score to individual units of the neural network in relation to a concrete output (classification or regression result). As with the Integrated Gradients method, a baseline is used: A neutral input is chosen (depending on the concrete use case) for which the activations of the individual units or neurons of the network are calculated. Thus, reference values are determined. Then the deviation - the 'score' - from these reference values is calculated for a concrete input per unit. The choice of the neutral input is critical and should be made using domain knowledge. In some cases, it makes sense to determine several neutral inputs and to calculate the individual scores based on several values.

Pros:
- Positive and negative influence of individual input values can be displayed separately
- Use baseline: intuitive approach
- Enables fast approximation for integrated gradients

Cons:
- Correct choice of baseline unclear → widely varying results

(Shrikumar et al. 2016; Shrikumar et al. 2017; Salehi 2020)

3.7.5 Guided Backpropagation und Deconvolution / DeconvNet

Explanation Type:
Decision explanations; Post hoc

Applicable to:
Neural networks, in particular convolutional neural networks; image data

Technical Background:
With Guided Backpropagation or DeconvNet (Decon-volution), important features of the input as well as individual layers of a neural network can be visualized. In both methods, the activation values of the individual units are mapped back to the respective input by the neural network in order to identify the input values that are decisive for a concrete classification with a saliency map. The same components are used as in a Convolutional Neural Network - e.g. pooling - but "in reverse". The process of traversing the network from back to front is also called backpropagation. The two methods Guided

Backpropagation and Deconvolution or DeconvNet differ only in the concrete calculations of the backpropagation steps.

Pros:
- Fast calculation, only one forward and one backward pass necessary
- Motivation behind the methods very intuitive
- Guided Backpropagation: "more selective" visualizations compared to DeconvNet

Cons:
- Strong focus on Convolutional Neural Networks → less suitable for other architectures

(Springenberg et al. 2014; Zeiler und Fergus 2013; Zeiler et al. 2011)

3.7.6 Activation Maximization

Explanation Type:
Model explanations; Post hoc

Applicable to:
Neural networks; focus on image data

Technical Background:
Activation maximization is used to gain knowledge about the structures learned by a neural network for recognizing different classes. The goal is to find input data that lead to a decision of the neural network that corresponds to a certain class with the highest possible confidence. Subsequently, the "perfect" input generated in this way can be checked for plausibility. With respect to the entire network, each individual unit can be considered and the activation of this unit can be maximized by a certain input. In this way, individual units and layers within the network can be examined and model explanations can be provided.

Pros:
- Explanations can be very fine-grained, e.g. for individual layers or units.
- Provides model explanations and insight into how the model works

Cons:
- Results are purely qualitative
- Interpretation difficult and very subjective (especially for low layers)

(Erhan et al. 2009; Ye 2020)

3.7.7 Sensitivity analysis

Explanation Type:
Decision explanations; Post hoc

Applicable to:
All models, regardless of their concrete implementation; image and text data as well as numerical data

Technical Background:
Sensitivity analysis is a concept that is applied across disciplines for the analysis of systems. In sensitivity analysis, individual input parameter values of a model are systematically varied (within the respective permissible range). These systematic variations, also called pertubations, can be used to determine which input parameters or features have the greatest influence on, for example, a classification result. Relevant features can be used as the basis for a corresponding explanation. Sensitivity analysis is model-agnostic and provides decision explanations in terms of feature importance in a very simple way. In one-dimensional sensitivity analysis, only one input value is varied at a time; in multidimensional variants, the influence of several varied input parameters can also be examined simultaneously.

Pros:
- Very fast and simple for differentiable models

Cons:
- Not suitable for non-differentiable models

(Cortez und Embrechts 2011; Baehrens et al. 2009)

Remark

Advantages and disadvantages were compiled based on the expert interviews, on information provided by the Bosch Center for Artificial Intelligence and the following additional sources: (Bhatt et al. 2019; Sundararajan et al. 2017; Google 2020; Gondal et al. 2017; Montavon et al. 2019; Shrikumar et al. 2017; Tjoa and Guan 2020).

4 THE CURRENT USE OF EXPLAINABLE AI IN INDUSTRY AND SCIENCE

4 THE CURRENT USE OF EXPLAINABLE AI IN INDUSTRY AND SCIENCE

This chapter presents the results of the survey conducted with representatives from AI-related companies and scientific institutions (number of participants: n = 209). The results reflect the statements of AI developers (about 75 percent of respondents) and AI users (about 25 percent of respondents) with regard to

- the use of specific data types, AI methods and models,
- the explainability of selected AI models,
- industry-specific requirements in terms of explainability as well as
- concrete target groups and implementation possibilities for explanations.

The explanation strategies presented in the previous chapter illustrate that the type of the available data must be taken into account for the selection of an appropriate tool. Some approaches are particularly suitable for the creation of explanations for image data processing AI systems, e.g. Grad-CAM or LRP, for other systems text-based knowledge bases might be needed.

When looking at the types of data that developers and users work with most frequently according to the survey (see Figure 3), it becomes clear that numerical data, image and video data, and text data play a particularly important role. Numerical data occupy first place in the survey and are used by about three quarters of the respondents.

Audio data is used much less frequently. Approximately ten percent of the respondents named additional data types; these include, for example, 3D, CAD and geodata and were summarized for the evaluation in the category "Other".

Under the hypothetical assumption that application problems would differ only in terms of data types (and application criticality, distribution of AI model use, and audience-related needs would not vary), the following could be concluded from the survey results. Either explanatory strategies should support different data types, or they should be particularly suited to one of these common data types.

In Chapter 3, however, it became clear that the use of certain AI models also limits the choice of explanatory strategies. For example, some of the approaches discussed, e.g. Integrated Gradients or DeepLIFT, are only applicable to neural networks as the underlying model type. Surrogate models or counterfactual explanations can be applied to neural networks as well as to other AI models.

The survey results on the use of AI models and methods show that the neural networks, which are often in the foreground in the public and expert debate, are by no means the only AI models currently used. Decision trees, which are easier to interpret, are used just as frequently.

Types of data used by respondents (multiple answers were possible)*.

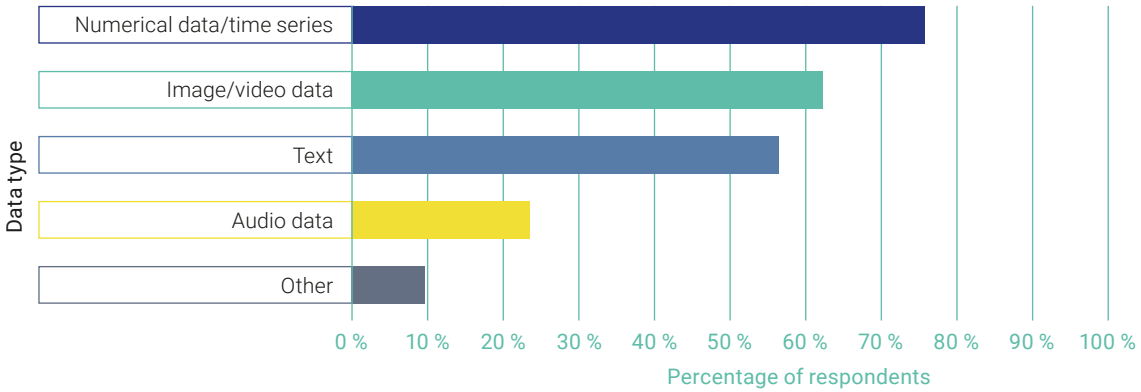


Figure 3 - Survey results: Numeric data is the most commonly used data type at approximately 75 percent.

* Some data types that could not be selected in the questionnaire, e.g. 3D, CAD or geodata, were indicated by several persons and classified in the figure under „Other".

Current and future use of selected models and methods according to respondents across all application industries (multiple answers possible)*.

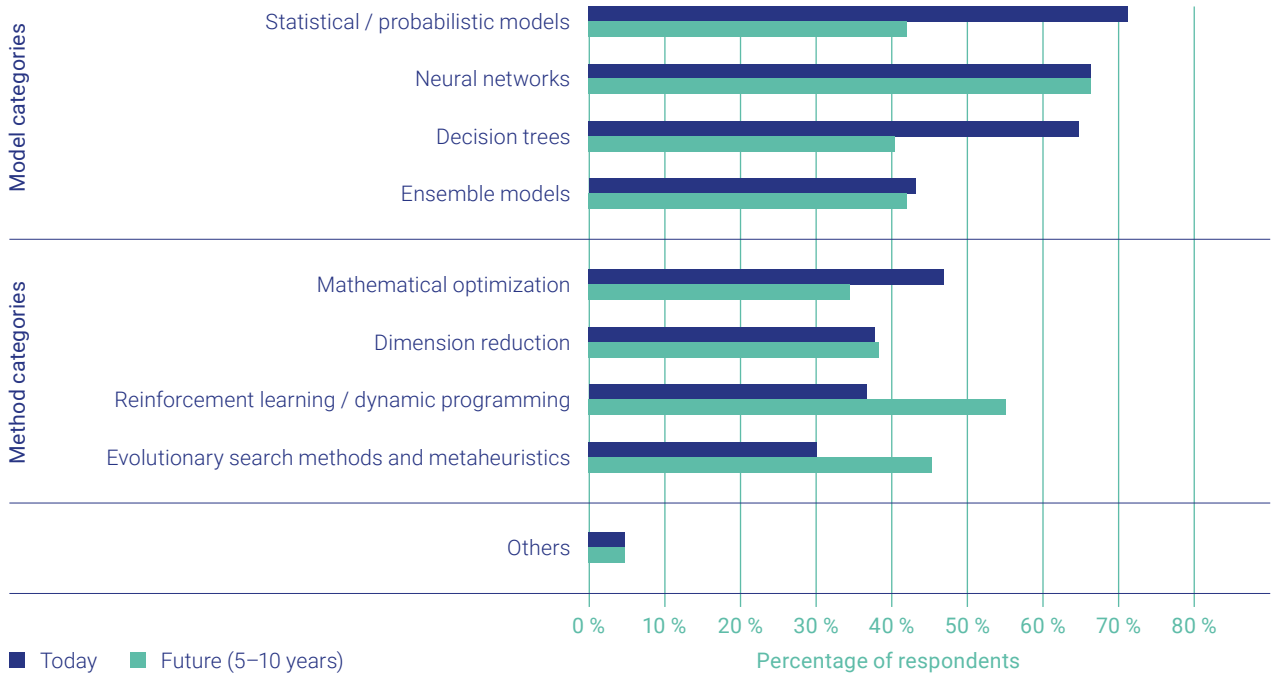


Figure 4 - Survey result: Neural networks will be the most important model type in the future, significant increase expected in the application area of reinforcement learning

* Only a limited number of categories of typical AI models and methods that were considered relevant for the subject of the study could be selected in the questionnaire. The graph shows the proportion of respondents who stated that they use or develop models or methods from the respective AI model or method categories. Several survey participants added the categories „expert systems", „knowledge graphs" and „semantic web" under „other".

Statistical and probabilistic models are currently the most widely used category of models. The combination or connection of several redundant models (ensemble models) is also a frequently used model category for a large group of respondents, see Figure 4.

Of the process categories that were available for selection, mathematical optimization approaches are the most widespread among the respondents. Methods for dimensionality reduction and reinforcement learning or dynamic programming are used only slightly less frequently. Evolutionary search methods and metaheuristics currently are the least popular among the considered approaches, but are nevertheless used by 30 percent of the respondents, which ultimately underscores a certain relevance of all method and model categories that could be selected here by the respondents.

A look into the future shows that the respondents see the strongest declining trend in two white-box model categories, namely statistical/probabilistic models and decision trees.

Since, according to the survey, the importance of neural networks will remain unchanged in the future (for about 66 percent of respondents), this could mean that a black-box model category will represent the most important model type in five to ten years. This suggests that explanatory strategies will also become increasingly important.

On the other hand, the survey shows that the importance of reinforcement learning and evolutionary search methods and metaheuristics will increase in the future according to the participants, and thus "on-the-job" learning and non-deterministic methods will be increasingly used. This study only partially deals with the "model type" of control policy in the context of a special use case (in section 5.2.2 of the following chapter). It should therefore only be briefly mentioned here that both process categories in this area of application can present potential challenges in terms of traceability and functional safety, depending on the methodological implementation and embedding in higher-level control processes. Possible "exploration phases" or non-deterministic modes of operation of autonomous systems often represent an exclusion criterion for approval, for example for control systems in the manufacturing industry (see also Section 5.2.3).

Assessment of the explainability of individual decisions (local explainability), which may have been increased by applying explanatory tools, related to selected AI models*.

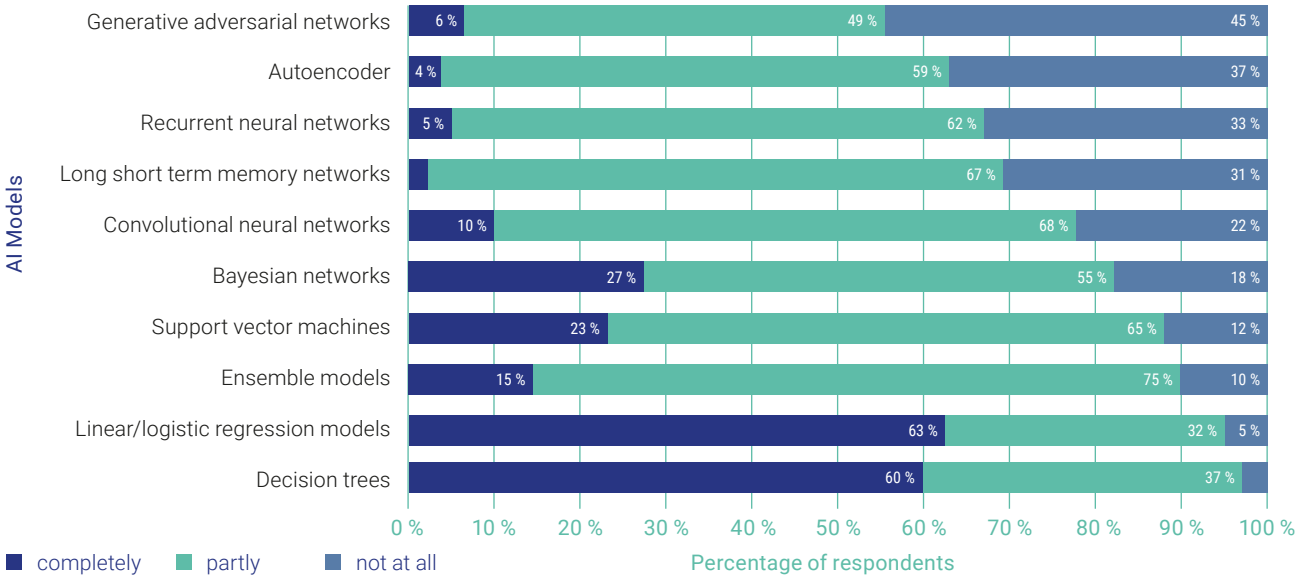


Figure 5 - Survey result: Providing decision explanations for neural networks is considered difficult.

* Persons only were asked about individual model types if they previously stated that they developed or applied models of the assigned supercategory. Respondents also had the option of stating „I cannot judge" as an assessment. In the figure, however, only information from persons who gave a corresponding judgement was taken into account. The individual methods were assessed accordingly by 16 to 81 persons and, in relative terms, by 50 to 84 percent of the persons surveyed in each case.

As part of the survey, participants were also asked to assess the explanatory power of individual decisions (local explanatory power) when using different AI models. The questionnaire explicitly stated that explanatory tools should also be taken into account where appropriate. What is striking about the survey results is that the five AI models that were rated as least explainable overall come exclusively from the neural network model family (see Figure 5). However, more than half of the respondents already consider these five AI models to be at least partially explainable locally, at least with the help of appropriate explanatory tools. A trend is emerging here that shows a discrepancy with the public debate, in which neural networks, for example, are often discussed as not being explainable at all.

On the other hand, it is also striking that a high proportion of the respondents consider the various model variants that can be assigned to neural networks to be not explainable at all. Here the range extends from just over 20 percent for convolutional neural networks to 45 percent for generative adversarial networks. This indicates that existing, relevant explanatory tools (Chapter 3) are not yet known to a significant proportion of respondents.

Overall, it can be seen that the theoretical division into white and black-box models, as often described in the literature (see section 2.2.3), is no longer equally reflected in the survey results when explanatory tools are explicitly taken into account. For a large proportion of the models, a majority indicated that they were "partially" explainable. Only decision trees and linear and logistic regression models were attributed by a majority (about 60 percent) to the "fully explainable" category. The survey also shows that Bayesian networks (nominal white-box models) received the label "not explainable at all" more often than support vector machines or ensemble models (both black-box models). This suggests that, in addition to the multilayeredness and the number of parameters of models, a certain experience in dealing with the models also plays a role.

The assessment of the current explainability of individual decisions (local explainability) of individual AI models contrasts with the concrete requirements from the individual industries. The survey results show clear differences between the industries (see Figure 6). Explanability is assigned a particularly important role for fields of application in which critical decisions are made. Explainability is often even mandatory in these cases,

e.g. for certification, seals of approval, standards, etc. In the healthcare industry, the creation of decision explanations is most important, which can be well plausibilized, since here wrong decisions can have fatal consequences in the worst case. Other industries where explainability of decisions is considered particularly important are finance, manufacturing, construction and process industries. The key role of local explainability results here from the requirements of customers and users, who generally would not accept a system that cannot explain individual decisions.

Production management and the process industry are characterized by a high degree of automation and special safety-related requirements; they are therefore demanding fields of application for AI applications in general and at the same time of great importance for Germany as a business location.

Depending on the respective industry, AI products and models are used by different people with different self-interests. Accordingly, the explanations generated must also be adapted to the respective target groups in order to offer added value. The addressees range from the groups of AI experts (developers) and domain experts (users) to internal or external auditors, management and possible end customers. In the healthcare sector, for example, patients represent the end customers. The AI

Importance of explainability of individual decisions (local explainability) according to application industries*.

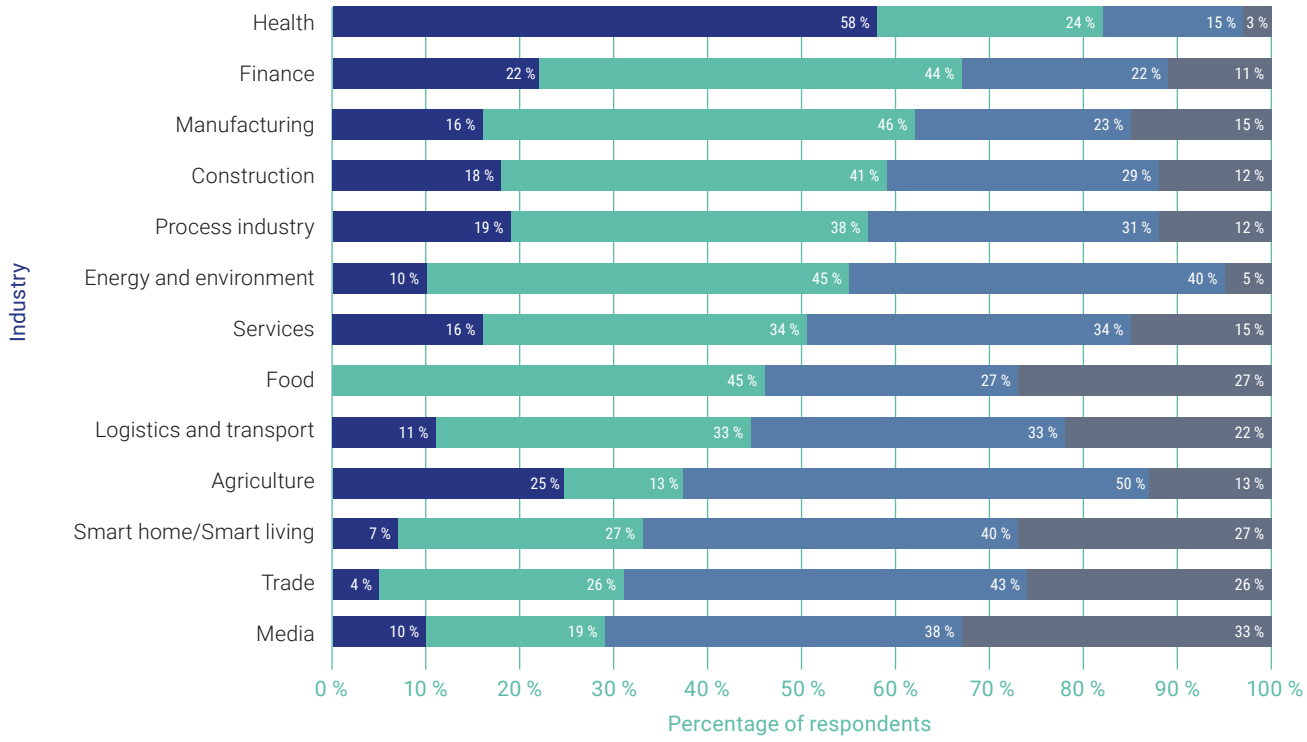


Figure 6 - Survey result: local explainability most required in sectors health, finance, manufacturing.

* Only people who had previously stated that they were using or developing AI systems in or for the respective application industry were asked about the individual application industries. In the figure, the industries were sorted according to the derived importance of decision explanations or local explainability (derived from the percentage of replies „mandatory" or „very desirable"). For other application sectors that few people specified as their target or application industry, local explainability was also classified as very desirable or mandatory several times, in particular for the rather unspecific application industry „IT / Software". For other individual additions such as "Legal Tech", "Human Resources" and "Public Security " local explainability was considered mandatory for the persons concerned.

systems - mostly decision support systems - are generally used here by medical staff, who in this context can be classified as a group of domain experts.

The survey results suggest that explainability is now particularly important for AI developers and domain experts (see Figure 7). This can also be observed when looking at all individual sectors. In addition, respondents expect the importance of explainability for most target groups to converge more and more in five to ten years: That is, explainability could generally also become more important for end customers, the management level, and internal and external auditors, while at the same time its importance will decline for AI experts and remain unchanged for domain experts.

The fact that the greatest change is predicted for end customers and external auditors can be interpreted in this way: Survey participants expect an increase in the importance of AI certification as well as in the general demand from customers for explainable AI as a result of a larger product offering. The expected loss of importance of decision explanations from the perspective of AI developers is somewhat more difficult to understand.

This trend could be driven by the two very contradictory assumptions of the survey participants, i.e. that the regulatory authorities will attach less importance to explainability in the future or that the use of black box models will be generally prohibited in certain areas of application ("high-risk") in the future. It is also conceivable that many scientific representatives, who rightly see themselves as AI developers, expect significant progress or a reduced need for research in the research field in five to ten years' time. However, if one excludes this slightly declining trend among AI developers from consideration, the growing importance of explainability becomes clear for all other stakeholders involved, right up to management level.

With the adaptation of the explanations to the corresponding target group, the question also arises as to the concrete implementation or presentation of the explanation that can bring the greatest benefit to the addressee. When asked in which way explanations can or should be implemented, most participants answered that graphic representations are well suited (see Figure 8). However, the survey, in which multiple responses were possible, also shows that there are basically many feasible ways to concretely design explanations; a universal solution cannot be identified. Rather, it can be assumed that the implementation depends on several factors - e.g. target groups or underlying data types - and that a well-suited solution must be found individually.

Assessment of the importance of decision explanations for target groups*.

Target group	Explainability of individual decisions (local explainability)			
	Today	Future (5–10 years)	Trend	
AI developers	76 %	56 %	▼	-20 %
Domain experts	59 %	59 %	▶	-1 %
Management	38 %	57 %	▲	19 %
End customers, end users	35 %	65 %	▲	31 %
Internal auditors	41 %	57 %	▲	16 %
External auditors	35 %	63 %	▲	28 %

Figure 7 - Survey result: Explainability today especially important for AI and domain experts, in the future a comparable importance is predicted for almost all target groups.

* The table shows how many percent of the survey participants consider the respective group to demand explanations of AI decisions (local explainability), evaluated across all application industries (while survey participants were asked with regard to their individual application industry or industries).

Types of explanations expected or desired by the participants (multiple answers were possible)*.

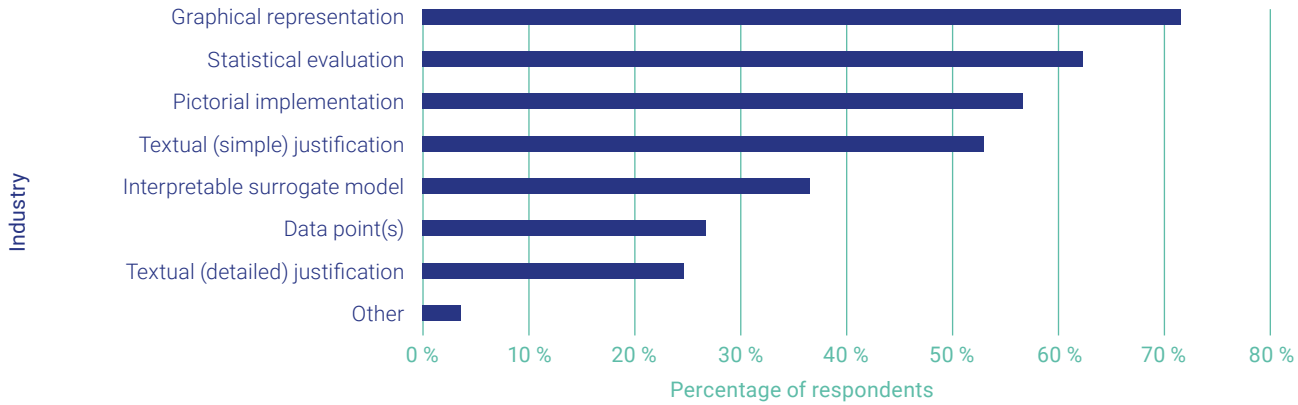


Figure 8: Survey result: Graphical representation considered most useful by survey participants.

* The graph shows the percentage of survey participants who expect the respective type of implementation or improvement of an explanation.

Explanation of the graphic:

- **Graphical representation:** e.g. display of the influence of individual characteristics on a decision
- **Statistical evaluation:** e.g. presentation of similar conditions and corresponding decisions
- **Pictorial implementation:** e.g. highlighting of decisive areas in image recognition
- **Textual (simple) justification:** e.g. naming the main reason for the decision.
- **Interpretable surrogate model:** e.g. generation of a decision tree for local approximation of more complex models
- **Data point(s):** e.g. data point(s) that would have caused a contrary decision (counterfactual explanations), representative data point for a particular class (prototype)
- **Textual (detailed) justification:** e.g. explanation of individual steps of the algorithm

5 USE CASES FOR EXPLAINABLE AI

5 USE CASES FOR EXPLAINABLE AI

The survey results on the industry-specific importance of explanations for individual decisions (local explainability) identify the healthcare sector relatively clearly as a field in which AI will not establish itself without sufficient explainability (see Figure 6 in Chapter 4). For a number of other industries, such as finance, manufacturing, construction, process industries, energy, and services, the majority of respondents also sees sufficient explainability as essential for AI to become established in the long term. However, while in these six industries the lack of acceptance of opaque AI modelsis mainly due to reservations among customers or users, in healthcare the regulatory hurdles for such systems are already seen as insurmountable, according to the survey participants.

Therefore, two use cases from the healthcare sector are presented below. This is followed by one use case from the general manufacturing industry and another use case from the process industry. These application areas are characterized by a high degree of automation and special safety requirements. They are therefore challenging fields of application for AI in general and at the same time of great importance for Germany as a business location. The information on the use cases was mainly derived from the interviews with experts and information from the literature[6].

5.1 Use cases in healthcare

In the medical field, AI algorithms are used to solve various problems. Image data like X-ray findings, digitized tissue sections or MRI scans, but also text data (medical reports and medical findings) or sensory data (electrocardiogram and blood pressure readings) can be analyzed more quickly and often more precisely with the help of AI algorithms. The evaluation of the analysis results can help to detect abnormalities in the data at an early stage and to react to them with appropriate further examination and treatment steps.

In addition to the important task of establishing acceptance and trust among users and those affected, healthcare-specific regulatory requirements must be met and privacy- and data security-related precautions must be taken in order to be allowed to include AI support in critical decisions. Especially the strict regulatory hurdles compared to other industries complicate or delay a broad availability of AI solutions on the market (BDVA Task Force 7 -Sub-group Healthcare 2020). After the following descriptions of two use cases - AI-supported image analysis and automatic analysis of medical reports - the regulatory aspects in the healthcare sector are discussed separately. The two use cases represent two typical applications of AI in the healthcare industry employing image processing and natural language processing (NLP)[7].

6 The team of authors generated the descriptions of the use cases, and is thus responsible for any supposed oversimplification or incorrect representation of details.

7 According to our survey, image and text data are used most frequently, along with numerical data (see Chapter 4).

5.1.1 Use case: AI-supported image analysis of histological tissue sections

Compared to non-invasive medical imaging, e.g. X-ray or Magnetic Resonance Imaging, histological imaging uses removed tissue samples that are stained with dyes or dye-labelled antibodies. Staining makes structures and cells visible. After the staining process, pathologists examine histopathological tissue sections of samples taken from a patient with suspected cancer (as part of a biopsy) for abnormalities. At present, most of this work is still carried out manually.

In corresponding research approaches[8], it is being tested how AI algorithms can support pathologists in their work. On the basis of digital scans of tissue sections, trained AI models can help to detect abnormalities and patterns that indicate a disease. The use of AI in this use case can help improve and speed up the analysis of medical image data by increasing the information gain for domain experts (Nagpal et al. 2018). The result of an AI analysis, for example, might reveal a wide variety of anomalies in the data. The pathologist can then combine these findings with his own experience and potentially make a more differentiated diagnosis, reducing the risk of missing critical anomalies[9].

The input data are image data, which typically do not have a uniform format and are characterized by a high resolution. Convolutional Neural Networks are primarily used for the analysis of the image data.

In contrast to white-box models, these black-box models are particularly suitable for automatic image processing. This is because complex patterns in images can be recognized implicitly without the developer having to specify rules. Therefore, AI-assisted image analysis of histological tissue sections, which will be considered in more detail below from the perspective of the relevant target groups, is also often based on CNNs, generated by supervised learning approaches[10].

In this use case, a physician is supported by an automated, AI-based anomaly detection system. The physician decides to what extent the results of the AI system should be taken into account for each individual diagnosis and thus also bears the responsibility. At the same time, the AI system assists with recommendations for potentially critical decisions that have a significant impact on the patient's physical well-being and can have fatal consequences in the event of misdiagnosis.

Target groups and overarching goals for the use of explainable AI

In the use case, the most important target group is the medical staff who will to use the system. At the same time, accredited bodies must be addressed (in this case the "Notified Bodies for Medical Devices"[11], see 5.1.3). Finally, the AI developers themselves also play an important role, especially with regard to the (further) development and improvement of the AI system.[12]

8 Project EMPAIA (https://www.empaia.org/) within the technology programme "AI Innovation Competition" of the German Federal Ministry for Economic Affairs and Climate Action.

9 An additional motivation for the use of AI is the stratification of patients into different treatment groups according to their individual risk. Based on this, individual therapy decisions can potentially can be derived, which, however, is no longer the focus of this use case. Precision medicine and personalized medicine can thus be pursued and improved in the future.

10 In principle, "Weakly Supervised Learning" and "Unsupervised Learning" as well as "Transfer Learning" can also be considered to improve models further, which is also being investigated in research.

11 Notified bodies are state-authorised bodies that carry out conformity assessments on behalf of manufacturers, e.g. for the approval of medical devices (Federal Institute for Drugs and Medical Devices, n. d.).

12 Use cases comparable to this one are often research projects, which is why corresponding researchers with AI or health expertise could in principle represent a further target group. In this and the following use cases, however, they will be consistently classified as developers if they develop corresponding systems. Due to the focus on practical implementations, researchers who investigate general medical contexts are not considered. Explanations could nevertheless help such researchers to discover associations in the data (such as correlations that give indications of causal relationships) and thus new biomarkers that can be used to improve the detection of diseases in general, or in this case for the diagnosis of cancer.

Another target group that could place more emphasis on the explainability of AI decisions in the future is, of course, the patients themselves. However, since the AI system is only used as a supporting analysis tool for domain experts (medical staff), it also remains the domain experts' task to provide personal explanations to patients. Therefore, patients are not explicitly considered as a target group in this use case.

A key motivation for using explainable AI in the use case is to "find" causal relationships (to relate particularities in the patient data to classifications of the AI system), which is especially important for domain experts - in this case, pathologists. Developers use the explanations to determine the confidence (functionality, robustness and stability) and thus identify fundamental vulnerabilities of the system with respect to various disturbances, such as statistical measurement errors, in input or training data. On the other hand, it is also the aim of developers to identify possible data bias, i.e. systematic error, in the training data (evaluate the fairness or detect data bias), as this can also result in the non-detection of tumours. In addition, regulatory or approval requirements must be met, but these are still largely unclear at present.

Explainability requirements from the perspective of the target group(s)

When using explainable AI in the use case considered here, the developers' particular aim is to increase confidence. This does not necessarily require explainability of the inner mechanisms of the model. However, it must be possible for this target group to identify if the data used to train the model contained a possible bias in order to minimize the risk of wrong decisions.

A mere explanation of how the algorithm works and how the given information is processed is not sufficient for this purpose (e.g., which layers of a neural network are responsible for recognizing the different structures of an image that supposedly shows a tumor). Rather, developers must be enabled to examine the factors that might influence the AI system's decision-making behavior, such as a possible bias caused by an unbalanced database. It must be avoided that the algorithm learns decision criteria during training that are not applicable in practical use or lead to erroneous results. This could be, for example, the classification of histopathological sections according to the scanner used or the date of acquisition. Another example is the use of image data which was already analyzed by medical experts. Pathologists may have already marked tumor areas manually - and the AI system then learns to recognize tumors only on the basis of these markings.

A bias in the training data could also result in a poor performance of the application for a different, for example significantly younger, patient group. The aim is to eliminate such error sources already during development.

The domain experts (healthcare professionals) who will use the explanations in practice should be able to check the system's decisions to see whether the results are reliable from a medical point of view. Even if concrete criteria for approval in terms of explainability have not yet been defined, several of the interviewed experts assume that at least individual decisions of the AI system on single patients must be comprehensible to physicians (local explainability). The lack of a detailed explainability of the inner model mechanisms would therefore not be a criterion for an exclusion of a system from approval. Rather, pathologists need to be presented with the indicators of why a specific decision was made so that they can use their own domain knowledge to make an informed assessment of the outcome of the AI system.

In particular, presentations of intermediate results can be helpful here, e.g., the segmentation of conspicuous image areas, in order to enable the domain experts to assess the plausibility of a given result.

Even if the regulatory requirements for explainability (see 5.1.3) are still very unclear, it is precisely these requirements that are crucial for the subsequent practical use of AI systems. Currently, a checklist of the "Notified Bodies for Medical Devices" is used for the approval of AI in the medical sector in Germany (Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland 2020). However, the checklist remains vague regarding explainability. One item on the checklist deals with the question of whether the developer has used explainable AI during implementation; another addresses the extent to which the end user has confidence in the product. Today, no concrete approval requirements can be specified with regard to explainability.

Explanation strategies

Post hoc methods already exist for explaining the decisions of neural networks, which focus in particular on the visualization of the results. In the input image, individual areas are highlighted according to their influence on the result of the AI algorithm. On this basis, the domain expert can decide whether the results are plausible, i.e. whether the highlighted image areas are so relevant that they can also be used as a basis for classification or segmentation from a medical point of view.

In the concrete use case of image analysis of histopathological tissue sections, neural networks are predominantly used to identify disease-relevant image areas indicating possible tumor cells. The focus of explanatory tools used here is on the generation of explanations through visualization. Specifically, the LRP method can be used to identify pixels in the input image that have a particularly high positive or negative influence on the classification result. LIME creates linear, local models to make it easier to understand individual decisions of the neural network. An advantage of LIME is the easy integration, LRP can provide explanations very fast. When using LIME in practice, it must be taken into account that the method is less suitable for very high-dimensional input data, so that the resolution may have to be scaled down. However, both methods should not be used without an understanding of how they work, and in this case only with prior knowledge of pathology and neural networks.

Potentially, other explanatory tools, such as Grad-CAM, Integrated Gradients, or DeepLIFT[13], intended for neural networks processing image data, are applicable to this use case. Alternative implementations of visualizations for the domain experts are possible as well.

Another approach is the use of Counterfactual Explanations. These can be applied to image data as well as other data types. The basic idea is to use counterfactuals to increase comprehensibility. Specifically, the target person is presented with hypothetical changes to the input data that would lead to a classification in a different class. In the example of image data, additional images are generated synthetically that look as similar as possible to the input images, but are each assigned to a different class (Goyal et al. 2019).

13 Pocevičiūtė et al. 2020 describe other concrete possibilities for the use of explanatory tools in pathology, for example Excitation Backprop, PatternNet or tSNE (Pocevičiūtė et al. 2020).

USE CASE
AI-supported image analysis of histological tissue sections at a glance

TYPE	Anomaly detection (decision support)
CRITICALITY	Very high (medical device)
DATA TYPES	Image data (2D, 3D), digitized tissue sections, high resolution
TYPICAL AI MODELS	Neural Networks (e. G. CNNs, GANs), (Explanatory deficits or black-box models)
(MAIN) GOAL GROUPS and respective overall objectives for the use of explainable AI	**DOMAIN EXPERTS** (medical staff): Check plausibility of causal relationships (find causal relationships)
	DEVELOPERS: Determine confidence (robustness, stability), test fairness (detect possible data bias)
	APPROVING AUTHORITIES ("Notified Bodies"): Verify compliance with approval requirements
CONCRETE REQUIREMENTS for explainability	**DOMAIN EXPERTS** (medical staff): assessment of the quality of individual ("local") decisions
	DEVELOPERS: Assessability of the model quality and revelation of bias in the training data (via "local" explanations of decisions)
	APPROVING AUTHORITIES: Verification of "comprehensibility" (concrete reulatory requirements currently under discussion), reduction of complexity
SUITABLE EXPLANATORY STRATEGIES	Decision explanations (post hoc), e.g. LRP, LIME

5.1.2 Use case: AI-supported text analysis of medical reports

A medical report summarises important data from a patient's medical history. The document can, for example, contain information on examinations performed and corresponding findings. Medical reports are used, among other things, for referrals to specialists or when patients are discharged from hospital, in order to pass on information from patients to specialist physicians or general practitioners.

When a differential diagnosis is made by the doctor, the complaints and symptoms recorded in the medical report are used as a basis. In the context of a differential diagnosis, clinical pictures that show similar symptoms are identified in a first step and then differentiated from each other with the aim of excluding irrelevant clinical pictures. In this way, a more reliable diagnosis can be made in many cases[14].

This process can be supported with the help of AI[15]. In concrete terms, Natural Language Processing (NLP) methods are used to automatically record and evaluate the content of a medical report. The aim of using AI processes is to support doctors by suggesting other possible clinical pictures whose symptoms correspond to those of the patient. The doctors are thus presented with a wider range of possibilities and the result of the AI system can be used like the second opinion of a colleague. Time saving is another advantage: NLP allows the doctor to automatically make a quick diagnosis.

It is not always easy to obtain an overview of the patients' medical history, which is often unstructured. In addition, the exchange between two physicians regarding similar patients is often difficult, since the identification of these patients is mainly done through personal conversations, which are very time-consuming. An automatic matching based on medical reports would be very helpful for this problem.

The AI models used in this medical text analysis use case are neural networks. Specifically, Transformer Networks are used, which have been repeatedly described in the literature as state of the art in NLP tasks (Otter et al. 2018; Wolf et al. 2019; Nambiar et al. 2020). Deep Learning, to which Transformer Networks are counted due to their many layers, offers the advantage over other methods that latent features are also detected: That is, information that is not immediately recognizable, which plays a role especially in the acquisition and processing of speech, such as indirect references or logical inferences. An indirect reference would be the description of the patient by words like "he" or "she" or "him" or "her". The networks are pre-trained on large medical datasets (Unsupervised Learning) and then adapted for the specific application (following the concept of Transfer Learning): Prediction of diagnosis through Supervised Learning. Only after training has been completed the AI system will be used in practice.

As with the previous use case of AI-supported image analysis of histological tissue sections, this is primarily an AI system for decision support. A similarity analysis is performed, which the physician can incorporate into the creation of the differential diagnosis. The responsibility lies with the medical professional. Likewise, the criticality is very high, as the results of the AI system are potentially used to make health-critical decisions.

14 If a patient is hospitalized with chest pain, for example, this symptom may indicate acute coronary syndrome or pulmonary embolism. In the context of a differential diagnosis, the aim is to identify these and other possible clinical pictures and then to exclude individual clinical pictures on this basis, for example on the basis of the patient's previous illnesses or risk factors (Strong Medicine 2018).

15 The article on differential diagnosis (https://www.BMWK.de/Redaktion/EN/Artikel/Digital-World/GAIA-X-Use-Cases/differential-diagnosis.html) describes practical examples and current challenges.

Target groups and overarching goals for the use of explainable AI

Domain experts, i.e. doctors, are to be supported by the AI system in assigning symptoms and complaints to different clinical pictures. The use of explainable AI should increase the information gain for the domain experts and thus make decision support possible in the first place. In addition, domain experts are driven by the goal of "finding" causal relationships, which can be achieved, for example, with the help of recognized medical publications that link certain symptoms to diseases and vice versa. Confidence - particularly important for developers - is to be increased by making the data basis more plausible through simplification, so that two patients with the same symptoms are also assigned the same clinical picture.

Explainability requirements from the perspective of the target group(s)

For domain experts and healthcare professionals, the central decision-making criteria of the system play a particularly important role. The physician must be able to decide for each patient individually to what extent a proposed clinical picture can be considered medically plausible on the basis of the symptoms present. In order to do this, a system that is intended to support this decision must be able to justify the content of individual decisions. In this way, a medical expert can efficiently assess whether a criterion that was decisive for the classification is either plausible or not medically reasonable. Consequently, this is the only way for the physician to decide whether the results of the AI system should be included in the differential diagnosis or not.

From the point of view of medical experts, it makes sense to provide appropriate explanations by displaying similar cases (patients with similar risk factors and corresponding symptom composition) or findings from the literature (publications from the medical field). Besides the fact that the use of medical literature as a data source per se adds enormous value to the AI system, the number and recognition of suitable sources can also be used as an indicator for the confidence of decisions. Moreover, even if this is a rather indirect consequence, the time-consuming exchange between physicians can be simplified by explainable AI, if the AI system makes the disease histories comparable in an automated way and accelerates the identification of patients with similar histories[16].

During the implementation of such an AI system, the developers are primarily concerned with the early detection of errors that arise due to medically implausible "features". This includes, above all, the consideration of the data basis. For example, the distribution and frequency of diagnoses should be examined (with the help of the medical experts) so that subsequent decisions by the system are not made on the basis of a bias in the database. This requires a systematic examination of the algorithm with regard to single variables and correlations of variables. For example, for an input variable such as age, it should be checked whether its systematic change or variation changes the prediction of the AI system as expected by the healthcare professionals or whether it is influenced by peculiarities in the training basis that do not correspond to reality. The focus here is on checking single-case decisions (local explainability), which can also be generated for AI models with black-box components using appropriate explanatory tools. The data basis for the AI model becomes more comprehensible primarily through plausibility checks.

As described above, specific requirements for explainable AI in the healthcare sector are still under discussion from the perspective of the regulatory authorities. In principle, it must be demonstrated to the regulatory authorities that the system achieves the objective pursued - e.g. increased efficiency or improved differential diagnosis. This point is also essential for other target groups such as hospital management, as indicators of economic efficiency can be derived from it. As in the first use case, a possible future target group is that of patients who are interested in an explanation of the diagnosis they have been given, but who are not explicitly considered here either[17].

16 On the basis of a corresponding highlighting of patient cases with high similarities, it could be decided in a timely manner whether a further exchange between the treating physicians appears to be useful.

17 The attending physician uses the AI to make a diagnosis, which is then communicated to the patient. The diagnosis is therefore only discussed directly with the attending physician; the patient does not have to be able to understand the result of the AI system independently.

Explanation strategies

In the considered use case, two concrete strategies are pursued to ensure explainability. Both focus in particular on the target group of domain experts (medical professionals). However, the provision of decision explanations is not ensured "post hoc" via the detour of an additional tool (as in the first medical use case). Instead, the model itself can provide decision explanations[18]. The basis for providing explanations are black-box models (neural networks), which are extended by prototypes or external knowledge bases, so that the model itself can provide medically comprehensible reasons for individual decisions and fulfils the requirement of local explainability.

In the first approach, the neural network works with prototypes. In the simplest case, a prototype is a single representative instance from the data basis. In this use case, a prototype would be a clinical picture with corresponding symptoms, derived from a medical report. For more generalization, multiple instances are often combined into a representative prototype (using supervised learning). For example, a patient suffering from influenza complains of a cold and headache and has a body temperature of 39° C. These symptoms are noted in the medical report. Another patient, who is also diagnosed with influenza, has a severe sore throat, cough, and also a fever. When creating the prototype for the clinical picture "flu", the symptoms of both patients would now be summarized and noted.

With the help of suitable prototypes, complex data sets can be presented to the users in a more comprehensible way. This explanatory approach differs from others (such as e.g. quantifying the influence of a parameter on a result or approximating the AI model in a post hoc manner) in that the individual prototypes do not only improve the comprehensability, but also determine the outcome of the given AI system.

If a "new" medical report is to be classified, the characteristics described (symptoms such as tiredness, cold, sore throat and increased temperature) are compared with those named in the individual prototypes and the prototype with the most matches is selected. Appropriate distance functions or classification methods such as K-Nearest-Neighbor can be used for this purpose. After the most similar prototype (in this example "flu") has been identified, it can be compared with the medical report and the crucial matches (cold, sore throat and increased temperature) can be highlighted to make the diagnosis of the AI algorithm comprehensible.

In the second approach, neural networks are combined with external knowledge bases, such as publications or general medical works in which diseases are described. The neural network learns the connection between symptoms and diseases on this data basis. Subsequently, the model can be further trained with medical reports. The knowledge bases can be understood as high-dimensional knowledge graphs in which the representations of thematically similar publications are placed close to each other. If a "new" medical report is to be classified, the AI model makes a decision that can be directly traced back to the statements of the publications from the knowledge base. This makes it easy to check conclusions and increases the comprehensibility of the model. When a decision is made, the physician is shown relevant publications for individual sections of the medical report that describe the facts under consideration - for example, specific clinical pictures that often occur with the symptoms described. In this way, the physician can draw conclusions about the credibility of the publication and the credibility of the algorithm's decision.

18 The literature contains contradictory information on the designation of such models. Some sources classify these models as antehoc (despite their BlackBox content) (Sokol and Flach 2019; Holz-inger 2018). However, there is some dispute in the literature as to whether ante-hoc explanatory power is a property that only white-box models may claim, or whether black-box models that provide certain explanations also provide them "ante hoc". In the following, ante-hoc explainability is only used for white-box models and explicitly referred to when black-box models are also meant, without using "post hoc" explainability tools for them.

USE CASE
AI-supported text analysis of medical reports at a glance

TYPE	Similarity analysis (decision support)
CRITICALITY	Very high (medical device)
DATA TYPES	Text data: Medical reports
TYPICAL AI MODELS	Neural Networks (Transformer Networks) *(Explanatory deficits or black-box models when used alone)*
(MAIN) TARGET GROUPS and respective overall objectives for the use of explainable AI	**DOMAIN EXPERTS** (medical staff): Increase information gain, check plausibility of causal relationships ("find" causal relationships)
	DEVELOPERS: Determine confidence (robustness, stability)
	APPROVING AUTHORITIES ("Notified Bodies"): Verify compliance with approval requirements
CONCRETE REQUIREMENTS for explainability	**DOMAIN EXPERTS** (medical staff): Enable decision support through substantive justifications (local explainability)
	DEVELOPERS: Deeper understanding of the functioning (through local explainability) → to improve the systems
	APPROVING AUTHORITIES: Verification of "comprehensibility" (concrete regulatory requirements currently under discussion), reduction of complexity
SUITABLE EXPLANATORY STRATEGIES	Decision explanations by prototypes and external knowledge bases in combination with neural networks

5.1.3 Regulation and certification in healthcare

Due primarily to criticality and data sensitivity, certification requirements for software systems in the medical field are particularly demanding. In recent years, various adjustments have been made in Germany regarding the permissible use of software components within medical devices. However, the formulation and implementation of concrete requirements for AI systems is currently still in progress.

From 26 May 2021, the new Medical Device Regulation (MDR) will come into force and replace the Medical Device Directive (MDD) currently in force (Remark: The editorial deadline of the original study was in April 2021). The MDD of the European Union was previously transferred into German law by the Medical Devices Act ("Medizinproduktegesetz"). Compared to the MDD, the MDR also contains some changes regarding the approval of software, which consequently also applies to AI applications. From now on, software will be classified in higher risk classes, resulting in more demanding requirements. The MDR describes four different risk classes, the classification of the applications is based on the intended use. Specific requirements are described, for example, for the development, validation, verification of functionality, production and monitoring of algorithms, which are in turn checked by the German state-authorised institutions ("notified bodies") that carry out the conformity assessments for the approval of medical devices.

However, there are no specific standards for the certification of AI systems, so it is often unclear how the specific requirements are to be implemented. In addition, there are numerous standards and guidelines that can potentially also be applied to AI systems. For the certification of AI systems, the Johner Institute - a well-known German company that offers consulting services for the approval of medical devices - has therefore developed a checklist as an orientation guide, which is used by the notified bodies as the basis for their own checklist. This explicitly restricts the approval of software in the medical field to pre-trained AI systems. Accordingly, no AI algorithm can currently be approved that continues to learn during operative use i.e. that changes its general behavior of decision-making due to a re-training with new data without official re-certification.

In the aforementioned checklist of the "notified bodies", the following requirements for an approval with reference to explainability are formulated (Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland 2020):

- The manufacturer should have analyzed the extent to which explainable AI approaches can make the developed model and/or its decisions more comprehensible.
- The applicability of different types of AI models should be investigated (especially comparisons with "simpler and interpretable" models should be made). With regard to the interaction with users, the extent to which they trust the system or want to review decisions should be examined.
- Further measures to be taken: Risk management (assessment of risks arising from the use of AI or its unintended use, e.g. with regard to input values or groups of patients), assessment of prediction quality, performance, training data sets (scope, origin, possible bias) and reproducibility of the results as well as creation of a post-market surveillance plan (activities or procedures to be followed when the product is on the market) to ensure model quality even after approval.

The "notified body" examines the aspects of the checklist individually, so that decisions are always made on a case-by-case basis and individual requirements can also be interpreted differently. From the perspective of system development, it is important to specify the approval requirements for an AI system in order to be able to address them appropriately. In particular, concrete requirements for explainability are unclear. However, corresponding specifications could support companies during the development process, accelerate certifications and increase the safety of patients. Explainability could also contribute to a possible certification of AI systems learning "on the job" in the future. Such AI systems that learn continuously offer the opportunity, for example, to improve diagnoses and prognoses through greater individualization. At the same time, however, the assessment of risks and the validation of safety aspects is significantly more difficult in this case (Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland 2020; Arbeitsgruppe Gesundheit, Medizintechnik, Pflege 2019).

Proposals for an approval of AI systems learning "on-the-job" have already been developed in the US. In a first draft paper, the U.S. Food and Drug Administration (FDA) describes a "Total Product Lifecycle Regulatory" approach, which includes a predefined plan for changes - including the type of change - after approval of the system. It must be possible to review these changes with regard to potential risks for patients. For example, newly added data must also be "quality assured" (compared to the original data and depending on specific application). In addition, it must be justified why the new data are necessary, and the specific goal of the new algorithm training must be explained. The previous system and the newly trained one should be compared with regard to previously defined prediction quality criteria, with the respective changes being transparent to the users. In the action plan published by the FDA in January 2021, it is described that the previously prepared draft paper will be revised according to the results of further discussions that have taken place. In addition, the creation of "transparency" for users of AI systems is to be pursued further and focused more strongly (U.S. Food & Drug Administration 2020, 2021; Working Group on Health, Medical Technology, Care 2019).

5.2 Use cases in manufacturing

The practical requirements of the manufacturing industry with regard to explainable AI reflect the fact that in this domain there is often a great deal of expert knowledge available on addressed processes, machines and plants and that economic efficiency plays the central role. Instead of processes in the human body, which AI systems in the healthcare industry analyze, the focus of software and AI system development here is on "human-designed" machines, plants or processes and their efficient operation.

In manufacturing, possible AI applications range from analysis tasks, such as machine monitoring or quality control, to planning support, e.g. for procurement processes, to autonomous systems, such as driverless transport robots or AI-supported process control. The entire spectrum of human-machine interaction is also a rapidly developing field of application.

Generally, a distinction is made between such manufacturing industries, where products are manufactured as countable units, and the process industry. The differences in the manufacturing processes result in differences in the risk potential. While potential accidents in the process industries, e.g., processing explosive and toxic materials, can have far-reaching consequences for people and society in the wider geographic area, potential accidents in other manufacturing industries (e.g. assembly of countable products) have more local effects on those affected and on the environment. Accordingly, the approval requirements for systems in the process industry are much stricter, which also has concrete implications for AI systems used in the domain and, in particular, the explainability of these AI systems.

As in the previous section, two use cases are presented. Section 5.2.1 describes a use case for AI-supported machine condition monitoring. It is one of the most typical AI applications that can be applied to many manufacturing industries. Section 5.2.2 presents a use case for AI-supported process control in which AI components are embedded in a larger overall system that controls safety-critical processes in chemical plants. While all components are under human supervision, the overall system is autonomous to a significant extent. The latter use case was selected because the AI-supported overall system must continue to learn during operation ("on-the-job-learning"). This is typically not the case with decision support systems that usually rely on pre-trained models.

While the creation of acceptance and trust by means of explainability is of great importance both in the process industry and in "discrete" manufacturing, the regulatory requirements that must be met for an approval in the respective fields of application can differ quite significantly. Section 5.2.3 discusses regulatory aspects of the manufacturing industry and corresponding distinctions.

5.2.1 Use case: AI-supported machine condition monitoring

Downtimes of individual devices, machines or systems can quickly result in enormously expensive production losses. This is especially true when they are embedded in complex production processes and are responsible for critical processing steps there. Effective early warning systems that indicate machine malfunctions or maintenance requirements can reduce or completely avoid possible downtimes and make maintenance management more economical.

Conceivable applications exist in all subsectors of the manufacturing industry. During machine or plant operation, large amounts of data are usually generated, which are usually time series of numerical type. In the most suitable case, corresponding systems also output explicit error codes if a machine failure has occurred or is imminent. Taking into account the time when an error code was sent, the recorded data packets can then be subsequently provided with corresponding "labels" indicating the occurrence or imminent occurrence of an error. If sufficient operating and fault data are available, this results in a classic application scenario for supervised learning methods, namely the detection of anomalies in machine behavior (condition monitoring), which are to be discovered as early and as reliably as possible.

This condition information can help domain experts enormously to plan maintenance measures either on the basis of their own expertise or on the basis of corresponding models, and to schedule the exact timing of the maintenance measure with knowledge of the machine condition (predictive maintenance)[19].

In principle, various appoaches employing black-box models, such as support vector machines, are well suited for the realization of the corresponding tasks of condition monitoring.

In the present case, however, the monitoring systems should output information about potential anomalies in machine behavior that is as comprehensible and localizable as possible. Statistical information on measurement and sensor data can be assumed as given. Therefore, the use case outlines an approach using ensembles of Bayesian networks on the one hand (approach 1) and a knowledge-based approach on the other (approach 2).

As in the use case of AI-supported image analysis of histological tissue sections, the aim in this use case is also to detect anomalies and provide decision-making support - however, here, in the manufacturing domain. Likewise, concrete classification results can often give an indication for a suitable maintenance planning. At the same time, the criticality of the application is high, since misclassifications or unrecognized signs of machine or plant damage can cause enormous economic damage.

Target groups and overarching goals for the use of explainable AI

The most important target group for explanations in this use case are the domain experts (maintenance or maintenance reliability teams), who are responsible for defining maintenance cycles and initiating maintenance processes. The core objective for the use of explainable AI in the use case is to make decisions qualitatively plausible for the domain experts, taking into account their understanding of the process ("finding" causal relationships). It is similarly important to convince this main target group, usually engineers by profession, of the statistical significance of individual decisions on an ongoing basis, e.g., of their robustness to varying measurement errors (determining confidence).

Finally, it may be a desirable goal to improve interaction possibilities for domain experts, so that they themselves can improve the explanations of AI systems or even the systems themselves by means of corresponding inputs.

19 For dedicated, AI-based planning of optimal schedules for maintenance measures (predictive maintenance), strictly speaking, it is necessary to have a degradation model available for each individual type of damage. On the basis of appropriate models, the remaining lifetime of a machine or plant can then be estimated using regression methods. However, since suitable models often cannot be set up due to a lack of corresponding damage cases, the term predictive maintenance is often used when actually only the detection of anomalies (condition monitoring) in the machine or system behavior is carried out.

This can be understood as a basic motivation of the developers of AI systems, while the concrete requirements in this regard are defined by the domain experts.

The developers of the AI systems can also be considered to be stakeholders, since - at least in the introductory phase of corresponding AI systems - they have to set threshold values that determine how pronounced an anomaly must be for the domain expert to receive a warning message. However, this motivation is equally close to the overarching goal pursued by domain experts (determining confidence), which is why the target group of AI developers is only marginally considered here.

Another objective not considered here, which may be relevant for developers, arises when machines or plants of similar design are to be analyzed (testing transferability) and fed into the same data pool for this purpose. Since in the present use case models are only trained individually and in relation to single plants, this aspect is not considered here.

Explainability requirements from the perspective of the target group(s)

In general, it can be assumed that domain experts must be convinced of the quality of the model, at least in the long term. Due to the corresponding responsibility for decisions that may have to be made under time pressure, the determination of explanations should not take too long; ideally, it should even be recognizable for the target group in which operating points the validity of the model is particularly good or particularly poor. For users with an affinity for AI, it can also be an important, or even essential, requirement to be able to understand in detail how a model was generated. Furthermore, confidence values provided should, if possible, also take into account statistical error distributions, if these are available (explainability of individual decisions). With regard to the target group, most likely engineers, and possibly given error distributions, statistical confidence indicators such as effect or signal strengths are particularly suitable.

Finally, an explainable AI system must enable intuitive and efficient interaction so that users can at least validate its behavior on a random basis or, desirably, even modify and improve the underlying model. The latter can be done in a minimal variant, e.g. by adaptable thresholds, or it can be realized by a designated human-in-the-loop approach. In the latter case, which involves a subject matter expert as a further “data source”, the explicit goal is to make the system more explainable on the basis of human-machine interaction and, at the same time, to allow it to continue learning through the input of experts.

An additional, at least desirable property of explainable AI is that domain experts can also receive alternative “explanation concepts” if required and preferred. A simple example would be that, when generating explanations for components, the individual explanation is based on their color instead of their supposedly complicated type designation.

Explanation strategies

If statistical confidence values and the consideration of additional statistical information are explicitly required according to the explanatory power requirements of domain experts, appropriate machine learning models must be used. Although suitable validation strategies (e.g. “cross-validation”) can be used to determine the predictive quality of each model on the basis of a given data set, supposedly available additional statistical information, such as occurrence probabilities of events, can only be comprehensively taken into account by certain types of models. If such additional information is available and if it is to be used explicitly to detect anomalies and to determine statistical confidence values in order to increase comprehensibility or at least plausibility, probabilistic models, e.g. Bayesian networks, are particularly suitable. Bayesian networks can be employed to represent probabilities of events and their interdependencies.

Convincing domain experts that AI models behave “correctly” is critical to the adoption of AI products designed to provide reliable condition monitoring of expensive machinery. For this reason, the target persons should be provided with the opportunity to perform a qualitative or quantitative model validation, e.g. via simulation, before deployment is considered.

However, it is advisable to let domain experts themselves choose respective scenarios during such a pre-validation, since “typical” scenarios chosen by the AI system provider could possibly be perceived as too selective. There are different ways to realize this in practice. A comparably simple possibility is to provide a suitable illustration of the input-output relationships of the original model.

In many cases, however, it is necessary to first make the basic, qualitative mechanisms of a model comprehensible to the domain experts by illustrating intermediate results. In these cases, it can be useful to provide the target group with a white-box surrogate model generated on the basis of evaluations of the original model and/or real operating data. In the simplest case, decision trees can be used for example. This surrogate approach has the advantage that domain experts can interpret decision boundaries and intermediate results more easily and quickly. Naturally, this is even more significant if the surrogate model is mechanistic, i.e., if it was created on the basis of physical laws anyway. Respective approaches via adaptable mechanistic surrogate models can be costly. However, model structures that allow for a derivation of suitable surrogate models might be available from the engineering process of the considered machines or plants (or similar ones). The parameter adaptation of the surrogate models to individual machines can then be carried out by means of curve fitting, i.e. parameter estimation.

The effort required to generate or adapt mechanistic models can pay off especially when machines and plants are very similar and process sequences are sufficiently easy to transfer. One example is turbo-compressor systems, which can vary enormously in size but, regardless of this, all function very similarly from a physical point of view and always consist of similar submodules. Another advantage of such mechanistic surrogate models is that cases of damage or operating data in unusual or potentially dangerous operating points can be determined by means of simulation if there is a lack of corresponding information or data. On the one hand, this can help to increase the process understanding of domain experts (helpful for training courses, etc.), on the other hand, additional data can be generated artificially, if necessary, for training the AI model (Bayesian networks in this case).

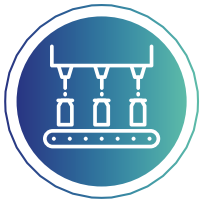
An alternative to meet the requirements of domain experts for model explainability even more comprehensively is to involve users directly in the generation of models and the associated explanatory approaches. Here, there are first promising approaches of corresponding machine learning methods on the basis of knowledge graphs. Currently, approaches are being tested[20] that combine inductive logical programming and reinforcement learning methods for the first time in order to obtain comprehensible machine learning methods that also include a so-called “human in the loop” concept. With appropriate methods, it should soon be possible for industry experts to interact with the AI system using natural language and thus - based on existing examples of machine anomalies and “normal” machine behavior - to define for themselves which explanations are suitable for them. In this way, transparent models can be generated, which can additionally be used to explain to the users of a corresponding condition monitoring system by means of natural language explanations why a system is in a permissible or impermissible state.

The most important task of the domain expert in this human-in-the-loop concept is actually to provide comprehensible designations for individual classes such as machine elements (e.g., “engine compartment”, “assembly line”) or tools (e.g., “Allen key”, “screw”). At the same time, with this approach, the domain expert is not necessarily needed to provide explanations (or to provide the actual condition monitoring functionality). If enough data is available in an appropriately processable form, processing can also be automated. However, automated assignment of class labels may limit the ability to generate natural language explanations that are clearly understandable to humans. Alternatively, these class labels could be obtained from additional external sources of information, if available.

20 RAKI project of the BMWK technology programme Smart Data Economy (https://raki-projekt.de/)

USE CASE
AI-supported machine condition monitoring at a glance

TYPE	Anomaly detection for the purpose of maintenance planning (decision support)
CRITICALITY	High (functional safety, economic efficiency)
DATA TYPES	Numerical data (sensor data, operational parameters or settings), text data (error codes, machine log data)
TYPICAL KI MODELS	**APPROACH 1:** Bayesian Networks **APPROACH 2:** Machine learning based on knowledge graphs
(MAIN) GOAL GROUPS and respective overall objectives for the use of explainable AI	**DOMAIN EXPERTS** (maintenance teams): check plausibility of causal relationships ("find" causal relationships); determine confidence (robustness, stability); improve interaction possibilities
	DEVELOPERS: Determine confidence (robustness, stability); improve interaction possibilities
CONCRETE REQUIREMENTS for explainability	**DOMAIN EXPERTS** and **DEVELOPERS:** Assessment (plausibility, statistical evaluation) of the quality of the models (concerns partial aspects of global explainability); assessment of the individual decision (local explainability)
SUITABLE EXPLANATION STRATEGIES	**APPROACH 1:** • Use and fitting of surrogate models (model plausibility), • Extraction of statistical quality parameters (Bayesian statistics) **APPROACH 2:** Natural language explanations (knowledge graphs)



5.2.2 Use case: AI-supported process control in the process industry

As early as the 1980s, methods were used in the chemical industry, for example, which can certainly be described as learning systems and whose advantages are only now being recognized in other industries. Examples include statistical filtering methods and model-predictive controllers. At the same time, the possible applications of data-driven AI methods in the process industry have only been investigated in greater depth in recent years. Industry experts expect that there is still untapped potential for increasing the efficiency of processes and plants, for example through the evaluation of image and video data or improved human-technology interaction.

In this context, the explainability of AI systems has a key role to play in enabling data-driven AI systems to be used in safety-critical application fields in the process industry, since both approval and acceptance depend on how well humans can comprehend and monitor the decisions made by AI systems.

To illustrate this, a use case of an AI-supported process control is described below. Data-driven AI methods are to be used in the context of "state detection" and for the determination of "optimal operational trajectories". The state determination (or detection) and the downstream control (determination of optimal operating sequences) are two closely related sub-applications that contribute equally to an AI-supported process control. In this context, state determination is generally a prerequisite for the determination of optimal operational trajectories. Only by a respective concatenation and execution in "real time" actual manipulated variables can be set and thus a process control be realized:

- *Condition detection via image data:*
  In the complex dynamic systems of the process industry, an essential subtask of process control is to obtain sufficient knowledge about the states of the system. Many of such states are often not directly measurable, but in the best case merely "observable" (i.e., roughly speaking, states can be estimated from the measurable variables if a sufficiently good model of the system is available). Elaborate sample analyses, for example to determine the exact proportional composition of material flows, are expensive and provide results with a time lag as well as with a much lower repetition rate than sensors. An inexpensive, yet rarely used, possibility to obtain additional data at high sampling rates is image data, which can be used, for example, to detect certain unwanted effects such as air bubble formation. In order to extraxt information from image and video data about state variables that are difficult to measure and which are subsequently used for process control, AI-based models and methods can be used.

- *Determination of optimal operational trajectories:*
  Finding optimal trajectories for steering plants in the process industry through start-up and shutdown processes is often a very challenging task. In addition to typically complex system dynamics, a large number of uncertainties, e.g. measurement errors or uncertainty regarding estimated quantities, as well as potentially critical constraints such as temperature or pressure limits must be explicitly taken into account. A particular challenge for robust model-based control are unforeseen, "discrete" events, such as switching operations.

In principle, an autonomous system can be assumed here, which has analysis, planning and control tasks. The criticality of the application must be rated as very high, since potential accidents can not only cause enormous economic damage, but also pose a risk for the health of local residents and the environment due to the processing of hazardous substances.

Target groups and overarching goals for the use of explainable AI

There are three target groups that must be considered when designing an explainable AI system for this use case. These are, first, the responsible authorities, without whose approval an AI-supported system cannot be used for process control. Secondly, there are the domain experts (operating personnel), who must be appropriately trained by the operator company in order to play their part in the safe operation of a plant. The third group is the developers responsible for the implementation of the overall process control. In this use case, these developers must also have sufficient domain expertise to be able to develop systems that are guaranteed, in accordance with the state of the art, not to violate relevant goals (health, environment, economic efficiency, etc.) during operation of the plant.

The most important overarching goal to apply explainable AI in the given use case is to be able to permanently check the susceptibility of the AI system to all possible kinds of malfunctions, especially in the operational phase (determine confidence), in order to be able to initiate emergency measures if necessary. In this context, it is essential that the domain experts with their individual understanding of the process are enabled to comprehend possible problems sufficiently comprehensively ("finding" causal relationships) and at the same time sufficiently quickly (increasing information gain/simplification) so that they can, if necessary, make quick and targeted adjustments on this basis (improving interaction possibilities).

Explainability requirements from the perspective of the target group(s)

For approving authorities, an AI-based component must be a part of a protection and emergency concept the moment it represents or influences a safety-relevant module. Since this is the case here, respective AI systems must be explainable with regard to the mechanisms of action, comprehensible with regard to risk minimization and assessable for testing with regard to the effectiveness of the safety concepts.

Even if the approval requirements for protection and emergency concepts do not specify any AI-specific requirements[21], approving authorities consequently require both decision and detailed and comprehensive model explanations (local and global explainability). Models and individual decisions must therefore be potentially verifiable - and thus explainable - for approval.

The operator company of a plant is responsible for safe operation. It must enable this safe operation by training the operating personnel and by suitable technical measures. Since the domain experts (operating personnel) thus have a central role in this safety concept, they must at least be provided with decision explanations (local explainability) in order to identify potentially safety-critical events and thus avert alleged accidents.

The task of the developers of process control is to develop systems that do not violate relevant protection goals during plant operation. As soon as AI-based processes touch upon safety aspects, local and global explainability are indispensable in order to integrate the AI modules in the overall system.

For this use case, however, another significant requirement arises. Since the real-time requirement must be understood as a general requirement of this automation-related use case, this naturally also transfers to explanations that must be made available to the corresponding stakeholders "in time" so that these can also act sufficiently fast if necessary (e.g. initiate safety measures or a shutdown of a subprocess).

Explanation strategies

How concrete explanatory strategies can be designed for this application is currently still being investigated. However, several promising approaches are already emerging.

21 Approval requirements are not specific to AI today in Germany. The Federal Imission Protection Act ("Bundes-Imissionsschutzgesetz"), the Federal Water Act ("Wasserhaushaltsgesetz") or the Hazardous Incident Ordinance ("Störfall-Verordnung") make no difference as to whether a (self-) learning system is used or not. However, mandatory requirements arise when an AI system directly or indirectly influences protection goals.

In order to meet the basic requirements of domain experts for decision explanations (local explainability), different approaches are considered fundamentally suitable for state detection via image data. For this purpose, post hoc explanation tools such as LIME, CAM and Guided Backpropagation have been preferred so far and their suitability for the use case has been analyzed accordingly. With these tools, interpretable explanations can be generated on the basis of image data for people with the corresponding prior knowledge, and biases can be detected comparably well. These tools are also considered to be sufficiently mature to meet the requirements of the domain experts for monitoring the automated state detection from images.

The requirement of very detailed comprehensibility of the decision-making processes, i.e. having to provide model and decision explanations (global and local explainability), predominantly concerns the second challenge in this use case: the determination of optimal operational trajectories. As explained above, this requirement arises from the perspective of the developers and the approving authorities. According to industry experts, there is every indication that the development of an AI system based solely on black-box models has no realistic chance of approval. The approval of any process control system requires the consideration of a comprehensive, approvable protection concept, for which the detailed comprehensibility of the algorithmic systems is considered indispensable (see also the following section 5.2.3).[22]

Consequently, the provision of classical post hoc explanations offered by analysis tools such as LIME is not considered sufficient by industry experts for the determination of optimal operating procedures.

In the context of the research project[23], which specifically addresses this use case, an implementation via the development of suitable, hybrid methods is aimed at. Simplified, within a corresponding hybrid AI, the white-box components could guarantee the fulfillment of the safety requirements, while the black-box components "supply" sufficiently verified information.

For example, an initially obvious approach would be to demand that only economically significant, but not safety-relevant manipulated variables be determined or calculated on the basis of black-box models. However, since in safety-critical systems all actuators interacting with the overall system must be regarded as potentially critical, it is fundamentally necessary to permanently (or at least regularly) check all effects of calculated manipulated variables internally for safety risks.

Potential safety risks that cannot be tackled by automated and always effective countermeasures must instead be adequately prevented in the operational phase by appropriate monitoring by domain experts (or developers) supported by a suitable human-machine interaction. In many situations, such interaction with human decision makers is quite feasible. In these cases, certain model adjustments that are economically promising but may pose a safety risk must be set to "pending" until they are verified by a human expert and/or simulation (e.g., significant variations of setpoint values recommended by the AI system). While such model adjustments may improve the operation of a process control system from an economic perspective (e.g., increase production throughput), the risk of undetected safety hazards (e.g., sudden undetected data bias that distorts state determination and thus poses an incalculable risk to process control) must be minimized. In certain scenarios, and with appropriate precautions in place, it is conceivable to delay the deployment of appropriate model adjustments until a human expert gives the approval. In the meantime, non-updated models could be used (as long as this does not jeopardize the protection goals), or traditional methods could be relied upon, e.g., state estimation using probabilistic filtering methods (e.g., Kalman filtering methods) to estimate non-measurable states.

However, given the degree of autonomy, it is necessary that domain experts are notified by the system when an appropriate decision has to be made. It is also conceivable that a course of action is suggested on the basis of past decisions.

With regard to the determination of optimal operating trajectories, the approach pursued here is to use AI-based methods to generate suitable "hybrid" models from plant and simulation data. The concept of a "hy-

22 In the event of liability-relevant accidents or incidents, the public prosecutor would examine whether the state of the art for the prevention of protection violations was taken into account, which in this field of application is oriented towards traditional and interpretable measurement and control technology as well as white box models. If necessary, significant economic consequences may arise for the responsible companies if a failure with regard to the protection concept is identified. In the event of an accident or incident that occurs despite operation in conformity with the approval and despite compliance with the protection concept relevant to the approval, the authorities granting the approval would be responsible.

23 http://keen-plattform.de/

brid" model means that white-box and black-box model components are combined in order to benefit both from the deterministics of mechanistic models and from the pattern recognition of data-driven approaches. From the resulting models, optimal control trajectories can be derived by means of modern control methods - in particular model predictive control.[24] These hybrid models have to be adapted and re-optimized during operation ("online") and monitored with regard to validity and performance in case of disturbances in the plant. Here, the explainability - both of the recommendations and of the adaptations through re-optimization - plays a major role for the trust in the resulting recommendation system.

Finally, industry experts expect that such a hybrid AI system must also meet the requirements of being able to provide suitable explanations "in time". This is because the core functionality of the overall system must also satisfy real-time requirements and the associated information content (of the explanation) must therefore also be able to be processed within certain time limits - either by a computer system, e.g., by means of simulation, but also by persons if required. In the latter case, system-based support of the responsible person is perspectively inevitable here, although the target group of domain experts is well acquainted with process control systems (or can be made familiar with them). The explanations must be aligned in terms of their level of detail to the expertise of the target persons and to the time constraints of the users and/or the overall process. By using the black-box components for somewhat less safety-critical and time-critical tasks, the significant demands on human "decision makers" and human-machine interaction in this use case can be reduced. As a result, it becomes possible for human decision-makers to react to events in the response times achievable for them and to make a major contribution to ensuring that an AI-supported process control behaves appropriately in terms of safety and possibly even learns incrementally.

24 Model-based variant of a reinforcement learning procedure that generates a control strategy or "control policy".

USE CASE
AI-supported process control in the process industry at a glance

TYPE	*There are two subtasks* (1) AI-assisted analysis (state detection via image data), (2) AI-supported feedback control (optimum operating procedures)
CRITICALITY	Very high (potential of accidents/incidents)
DATA TYPES	Numerical data (sensor data, operating parameters), image data

TYPICAL AI MODELS	For (1): Neural networks
	(explainability deficits / "black box")
	
	For (2): reinforcement learning (model predictive control) based on hybrid models
	(at least largely explainable)

(MAIN) GOAL GROUPS and respective overall objectives for the use of explainable AI	**DOMAIN EXPERTS** (operating personnel) and **DEVELOPERS** (process management): Determine confidence (robustness, stability, vulnerability); check plausibility of causal relationships ("find" causal relationships); increase information gain (simplification); improve interaction possibilities (especially for domain experts).
	APPROVING AUTHORITIES: Verification of "comprehensibility" and protection concept
CONCRETE REQUIREMENTS for explainability	**DOMAIN EXPERTS** (operating personnel): Explainability of individual decisions (local explainability)
	APPROVING AUTHORITIES and **DEVELOPERS** (process control): Single decision explanations and model explanations (local and global explainability)

SUITABLE EXPLANATION	For (1): post hoc explanations, e.g. LIME
	
	For (2): Integration of the black-box model components by means of hybrid modelling

5.2.3 Regulation and certification in the manufacturing industry

In the manufacturing industry, it can be observed that AI applications are increasingly being integrated into robotic systems or production plants. In particular, computer vision or AI-supported data analysis approaches are finding their way into the manufacturing industry to complement the algorithmic components that previously operated on a more rule-based basis.

The Machinery Directive 2006/42/EC of[25] of 17 May 2006 is the central standard of the European Union (EU) for the CE certification of such production machinery and systems. It regulates compliance with the principles for the safety of technical systems and thus provides the framework for approval - initially irrespective of whether or not AI elements influence the behaviour of the machine. The Machinery Directive thus provides a framework for detailed technical rules. For example, its Annex I contains the general health and safety requirements to be observed in risk assessment and reduction. The central objective is to define the functional design of the machine for the area of application and for its entire service life, so that persons are not endangered. This includes, for example, specifications for handling, control and maintenance of the respective system. Particular attention is paid to protective measures against mechanical hazards. In addition, the directive stipulates which information material on the machine and the associated protective measures must be available and in what form.

Up to now, the Machinery Directive has remained unaltered despite the technological development and the fundamental safety and usage principles of mechatronic systems based on defined, deterministic control and regulation concepts have endured. (Remark: the editorial deadline of the original German version of the study predated the proposal of the EU commission for a regulatory framework on machinery products published on april 21st 2021). However, a systematic evaluation of the directive in 2018 conducted by the European Commission indicated a need for future adaptations.

A survey of numerous stakeholders from the mechanical and plant engineering sector is part of this evaluation. It showed that the increased use of IoT[26] and AI in mechanical engineering products is expected to lead to a paradigm shift towards interconnected, autonomously deciding and even learning products of mechanical engineering. (European Commission 2018). In the case of using black-box models, the behaviour of the systems cannot be predicted with sufficient certainty. By embedding such models in technical systems, decisions of AI systems affect physical actions, even if only partial functionality is supported by AI. For example, an AI module for image recognition can contribute to navigate a robotic arm. However, misclassifications of the AI could theoretically result in incalculable safety risks, e.g. for human operators, for the processing of hazardous substances or for the surrounding infrastructure.

If, however, AI models in the manufacturing environment continue to develop during operation ("training on the job"), it may be extremely difficult or impossible to comprehend or trace decisions to a sufficient extent. If an AI product may pose a potential hazard to persons and no alternative safety precautions are (or can be) taken, certification according to the requirements of the current Machinery Directive is currently only conceivable if dedicated explanation strategies or tools are used. From a regulatory perspective, however, it is still largely unclear how to deal with AI systems that are (permanently) retrained in safety-relevant applications.

AI black-box systems that have already been pre-trained can also pose an obstacle to certification, especially if they influence physical actions or other safety-relevant functions of the system. If such influences are not comprehensively controllable and traceable, a risk for the safety of the overall system remains. Particularly when interacting with humans, such AI-supported systems

25 Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, available online at: https://eur-lex. europa.eu/eli/dir/2006/42/oj

26 IoT = Internet of Things

may therefore pose special safety risks that are not addressed by the previous certification requirements of the Machinery Directive.

It is still under discussion whether an adaptation of the Machinery Directive itself is necessary, which could be approached either by including ethical rules for autonomous systems ("robot laws") or by expanding safety regulations. Alternatively, technical standards may also be sufficient. However, the basic principle is that machines "should not enter into an uncontrolled state that would pose a danger to the operator or to uninvolved third parties". In addition to general safety devices and measures, explainable AI algorithms or corresponding warning systems can also make a significant contribution to facilitating human supervision and thus significantly reducing the aforementioned risks. It is therefore quite conceivable that future directives for the approval of learning systems take approaches into account that make use of explainable AI models and methods. However, for regulation-compliant system development and subsequent conformity assessment, it would be advantageous if the requirements for the notifications and explanations of the machine operators were specified as precisely as possible.

In the process industry, where hazardous substances are often processed under high pressure levels, additional regulatory requirements beyond the Machinery Directive must be taken into account (VERBAND DER CHEMISCHEN INDUSTRIE e.V. 2012). The applicable Hazardous Incident Ordinance[27] (SEVESOIII Directive) of the Federal Immission Control Act consequently prescribes high standards for the certification of process safety. Production concepts, for example, must be fundamentally comprehensible. Furthermore, it must be verifiable that the actual implementation corresponds to the planned concept. For the use of AI in this context, it must be ensured that the decisions of the overall system are sufficiently transparent, repeatable, comprehensible in detail and correctable if necessary.

The German Institute for Standardization (DIN) has been pursuing a dedicated AI roadmap for standardization since 2020 (Wahlster and Winterhalter 2020). Industrial automation is a key topic. In addition to software standardization for industrial applications, the requirements for learning technical systems are also considered. On the identified challenge of "explainability and validation", the VDE has already developed the technical Rule E VDE-AR-E 2842-61-1:2020-07[28] was published. It describes the terminology and basic concepts of explainable AI. Building on this, quality criteria and reproducible, standardised test procedures for reliable AI systems are to be developed in the national implementation programme "Trusted AI". It is not yet possible to predict when they will be applied to technical systems (Wahlster and Winterhalter 2020).

As a result, the same safety regulations currently apply to AI-supported systems as to conventionally controlled products, even though initial standardization efforts are underway. This means that there are liability regulations for the manufacturer that correspond to the product liability directive.[29] In particular, there are no recognized processes for the certification of AI-supported systems. This is especially true for systems that significantly change their behavior during operation without this change being subject to human supervision, so regulations currently exclude learning systems of this type. Even pre-trained AI systems that do not evolve may only be used under controllable conditions. Explainable AI algorithms are thus a basis for implementing learning and decision-making processes of AI systems in a comprehensible and thus controllable manner and for significantly expanding the spectrum of certified applications for technical systems.

27 Twelfth Ordinance on the Implementation of the Federal Immission Control Act, available online at: http://www.gesetze-im-internet.de/ bimschv_12_2000/index.html

28 VDE Application rule E VDE-AR-E 2842-61-1:2020-07, Development and trustworthiness of autonomous/cognitive systems - Part 61-1: Terminology and basic concepts, https://www.vde-verlag.de/ norms/1800574/e-vde-ar-e-2842-61-1-application-rule-2020-07.html

29 Law on Liability for Defective Products/Product Liability Directive, available online at: http://www.gesetze-im-internet.de/prodhaftg/ index.html

5.3 Overall consideration of the use cases

When comparing the four use cases, it is first noticeable that two general motivations for the use of explainable AI are considered important in all four cases, namely to check the plausibility of causal relationships ("find" causal relationships) and to determine confidence. At the same time, however, there are very individual motivations for the use of explainable AI: e.g., increasing the actual information gain through explainable AI or improving the interaction between humans and AI systems:

- In the use case for image analysis of histological tissue sections, the general motivation to "find" causal relationships is clearly the top priority for the domain experts (pathologists). In this case, the medical experts want to see at a glance why a specific classification decision was made with regard to a tumor detection (local explainability). In the best case, the pathologists can then trace the intermediate steps that led to the decision. Subsequently, the experts can decide whether the AI output should be taken into account in the medical diagnosis or discarded. With regard to approval processes, experts also assume that at least individual decisions on individual patients must be comprehensible for physicians[30]. A solution with post hoc explanation tools such as LRP or LIME visually highlighting image areas for the domain experts are pursued here as an explanation strategy. These post hoc explanation tools are also used by AI developers whose primary goal is to test fairness (detect data bias) or to determine confidence.

- The overarching goal to verify the plausibility of identified relationships ("find" causal relationships) is also the core motivation for the use case for machine condition monitoring: Domain experts (in this case engineers) first want to be able to investigate the discovery of possible anomalies by an AI system before they initiate any maintenance measures. An absolute minimum requirement in this case is therefore the provision of explanations for individual decisions (local explainability). Since typically no official approval is required for such AI systems, there are also no regulatory requirements on how explanations should be designed. However, maintenance experts often have to make decisions of great importance (in terms of safety and economic efficiency) under time pressure. Therefore, the explainability of model action mechanisms (global explainability), which enables pre-assessment of model reliability by domain experts, is usually critical to whether an AI system is ultimately used for condition monitoring in practice. Two different approaches are pursued as explanatory strategies for the use case. On the one hand, Bayesian networks and a surrogate model are used in order to provide users with intrinsic statements about the probability of events occurring and an illustrative model that can be simulated flexibly. A second approach is based on a machine learning methodology based on knowledge graphs. Natural language explanations are provided, which the user her- or himself can adapt to the individual requirements. The interaction between humans and AI systems is improved by this approach, which can also be a general motivation for users to use explainable AI.

- The overarching goal of increasing the information gain of the domain experts is the central motivation for the use case of medical text analysis of medical reports. In this context, the concrete, case-related indications as to why a particularly close proximity between patients' disease progressions was detected are indispensable as key information for the medical staff. This is the only way for medical experts to efficiently assess whether a criterion that was decisive for the classification of the AI is either plausible or not medically meaningful. For this purpose, a system that is to support this decision must be able to justify individual case decisions in terms of content (local

30 The first approved products for AI-supported radiological image analysis exist on the market, which gives an indication that the explainability of individual decisions (local explainability) was already sufficient for approval, at least in individual cases.

explainability). The basis for providing explanations are nominal black-box models (neural networks), which are supplemented by prototypes or external knowledge collections. Consequently, the resulting model itself can provide medically comprehensible reasons for individual decisions by visually highlighting relevant passages in medical reports or external publications for the target group.

- In the use case of AI-supported process control, there are different motivations for the use of explainable AI. However, the overarching goal of determining appropriate confidences is particularly crucial, especially with regard to the effects of individual decisions for the complex overall system. Undetected errors in the visual state recognition or susceptibility to disturbances and bias in the "hybrid" models can, in case of doubt, mean incalculable risks for the robust and stable control of the chemical plants. Therefore, this use case also reveals the most far-reaching explainability requirements compared to the others. Beyond the explainability of individual decisions (local explainability), the detailed explainability of model mechanisms (global explainability) is also required here. The approach taken is to create suitable "hybrid" models from mechanistic models and simulation data, as well as image and sensor data, which combine white-box with black-box components to create self-explanatory plant models. Advanced process control approaches could then take advantage of the plant models improved by machine learning for the "on the-job" generation of time- or energy-optimized operational sequences.

Of all the use cases considered, the provision of model explanations (global explainability) is a strict approval requirement only for AI-supported process control. Although technically experienced persons are available to supervise a corresponding AI system, it is impossible for them to check every individual action of the overall system or process control. Instead, responsible individuals must be actively made aware that decisions are to be made by them. In particular, if safety-relevant user decisions are not taken (in time) by the specialist personnel, the system must independently initiate alternative measures in accordance with a protection concept to be defined.

Requirements for the form and scope of explanations, as well as for the amount of time that a generation of explanations may take, are highly application-specific. Explanations must be adapted to the expertise of the target persons and to the time constraints of the users or of the process as a whole.

Requirements for the form and scope of explanations, as well as for the amount of time that a generation of explanations may take, are highly application-specific. Explanations must be adapted to the expertise of the target persons and to the time constraints of the users or of the process as a whole.

This requires explanations that must be available at the same time or at least shortly after the actual decision or recommendation of the AI system and must also meet the - sometimes conflicting - requirements for explanations: simple, brief and comprehensive.

Further details and references on the approaches used in the use cases can be found in Chapter 3 and in the Glossary in the Appendix A. Excluded from this is the Machine-learning approach based on knowledge graphs and the hybrid modelling approach, both of which are still too much in the research stage to be discussed in detail in this study.

6 PRACTICAL FIRST STEPS: ORIENTATION GUIDE FOR SELECTION OF EXPLANATORY STRATEGIES

6 PRACTICAL FIRST STEPS: ORIENTATION GUIDE FOR SELECTION OF EXPLANATORY STRATEGIES

In the following, recommendations for the selection of explanatory tools are presented, which were derived from the interviews with experts, literature research, the results of the benchmark tests from the study conducted by the "KI-Fortschrittszentrum 'Lernende Systeme und Kognitive Robotik'" of Fraunhofer IPA (Schaaf et al. 2021), and information provided by the Bosch Center for Artificial Intelligence. An overview of the tools and the underlying advantages and disadvantages can be found in Chapter 3.

Before selecting a suitable explanation tool, the design criteria for the target system must be taken into account. These considerations should include, in particular, the target groups of the explanation, the types of underlying data, and the selected AI model that makes the decisions.

With regard to the selection of the AI model, it should (in general) always be investigated whether it is possible to use a less complex, thus more comprehensible model for the solution of the initial problem that still meets the specific requirements. Regarding transparency, ideally, a white-box model is used. This is mainly because post-hoc explanations for decisions of black-box models can be problematic, as they try to simplify how the model works, but cannot represent it in its completeness. It follows that these explanations are not always completely accurate and are rather approximations that can also lead to errors in interpretation (Rudin 2019).

With regard to the selection of the AI model, it should (in general) always be investigated whether it is possible to use a less complex, thus more comprehensible model for the solution of the initial problem that still meets the specific requirements.

White-box models and AI models that themselves provide both decision and explanation offer the advantage that no additional analysis tool or surrogate model is needed to explain decisions. Rather, the original model itself is self-explanatory or it provides explanations along with the decision. Hence, there is no need to deal with the functioning of an additional analysis tool responsible for the explanation. The employment of white-box models, additionally, can enable the user to achieve a deeper understanding of the AI algorithm itself.

The majority of the interviewed experts considered most of the post hoc explanation tools (discussed in Chapter 3) only partially suitable for providing individual explanations of decisions for AI users, e.g. domain experts. The tools have the disadvantage that their handling is not intuitive for users, so that the correct interpretation is not automatically assured. In general, several experts expressed in the interviews conducted for the study that intuitive explanatory strategies must also be made available for users without AI expertise. Counterfactual explanations are a good example of this. Similarly, many experts considered surrogate models to meet the requirements of providing intuitive explanations - despite the inevitable discrepancy between the initial model and the surrogate model.

For example, the decision criteria can be read directly from the decision trees frequently used as surrogate model. The use of prototypes is also advisable in this respect in order to provide content-related explanations for decisions, as this approach also makes decisions plausible in a comparably intuitive form. Nevertheless, corresponding explanation strategies not only

make decisions of AI systems more comprehensible for domain experts. Also, AI developers can benefit from their use and gain new insights into the decision-making process of the model in order to improve the quality of the AI systems.

For AI developers, both the generation of decision explanations and model explanations can play an important role. Some experts see the discussed post hoc methods - with the exception of counterfactual explanations - as mainly suitable to support AI developers in improving algorithms. When using black-box AI, the methods Integrated Gradients (for neural networks) and SHAP (model-agnostic) are currently particularly well suited. If a fast approximation is sufficient, DeepLIFT can be used instead of Integrated Gradients. For the application of SHAP on multi-layered models with a high number of parameters or for the processing of high-dimensional data, it must be examined whether the runtime is still acceptable. In general, it should be noted that the usability of the methods strongly depends on the respective use case and it should be examined individually to what extent they meet the individual requirements. Assessment criteria of individual methods are often subjective and due to user preferences. However, when selecting a specific method, its functionality and disadvantages should be well known. As a further recommendation, it was noted that AI developers should not rely on one method only. It is advisable to test several methods in order to be able to recognize and circumvent problems of a method that may not be immediately apparent for the specific application at an early stage.

In the study conducted by "KI-Fortschrittszentrum 'Lernende Systeme und Kognitive Robotik'" of Fraunhofer IPA (Schaaf et al. 2021), various methods were investigated e.g. in terms of runtime performance and fidelity of the explanations[31]. For application to image data, the Integrated Gradients and LIME methods were rated best because of their fidelity to the model. SHAP achieved less good results in this regard. However, the two first mentioned approaches need more time to generate explanations than other methods, for example LRP. This result shows a difference to the assessments of the experts, who rated Integrated Gradients (in terms of runtime) as well suited for image data.

When applied to tabular data, LIME and SHAP achieved very similar results. Counterfactual explanations stand out because model fidelity is always given with this approach. However, the generation of explanations takes a relatively long time. The surrogate model (here: a generated decision tree) also performed well - particularly with regard to runtime performance (Schaaf et al. 2021).

To support the selection of a suitable explanatory tool, the described recommendations were summarized as an "orientation tree" (see Figure 9). When using it, it should be noted that only a selection of already well-established explanatory strategies and tools was considered and that the information is also based on experience with concrete use cases. The "orientation tree" shown is intended to represent the findings obtained in the course of the study in a simplified manner and to provide rough orientation when selecting explanatory strategies and tools.

31 Fidelity (of reproduction) indicates the extent to which the explanation reflects the actual behavior of the model.

The orientation tree

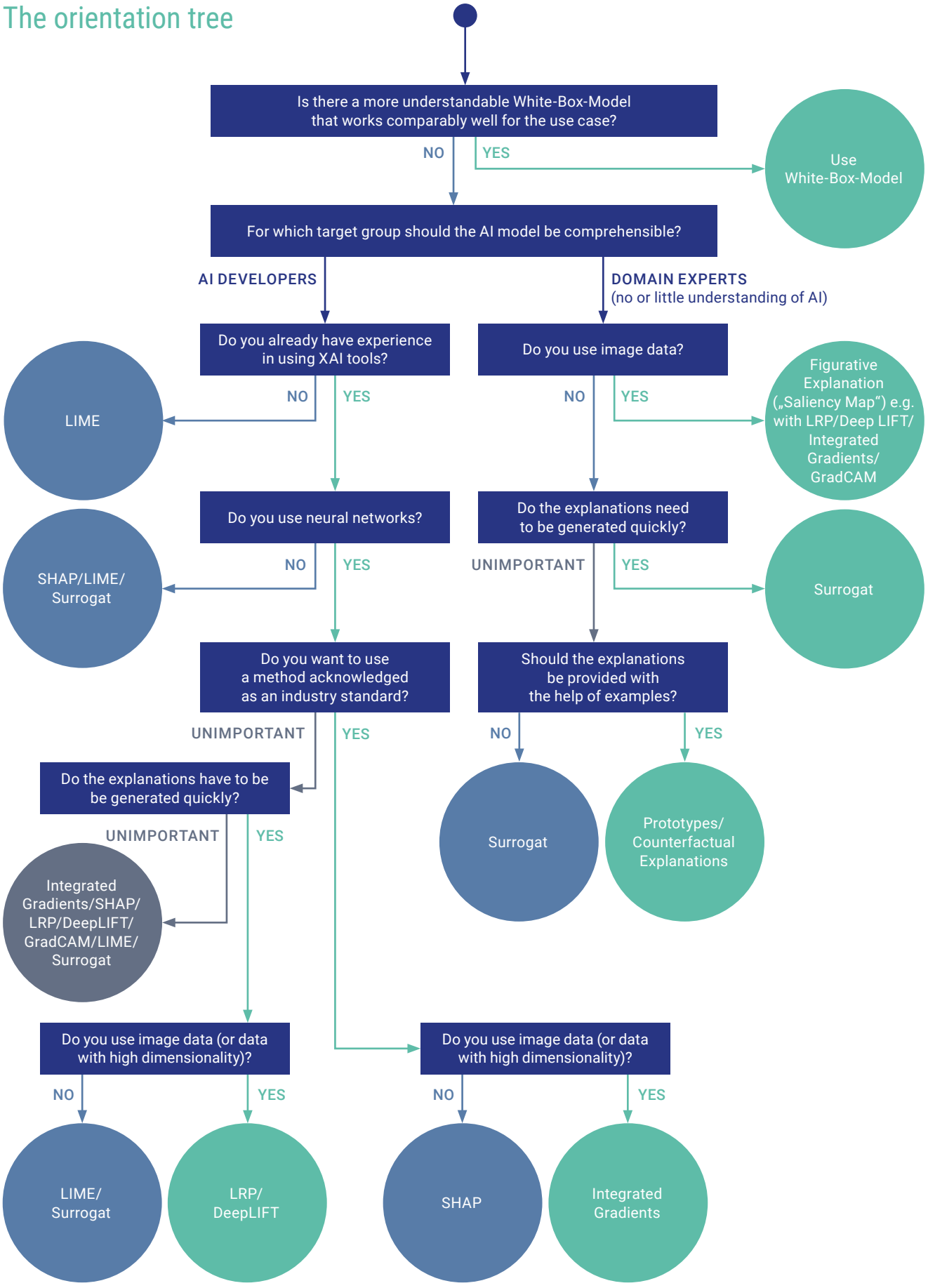


Figure 9: The orientation tree supports the selection of suitable explanation tools (XAI tools)

7 CHALLENGES AND NEEDS FOR ACTION FOR THE ESTABLISHMENT OF EXPLAINABLE KI

7 CHALLENGES AND NEEDS FOR ACTION FOR THE ESTABLISHMENT OF EXPLAINABLE KI

In the previous chapters, technical possibilities were presented for ensuring the explainability of AI systems for concrete use cases. In the interviews with experts conducted for the study, the future technical and regulatory challenges and needs for action for the realization of explainable AI systems were also discussed.

The experts were asked for their assessments of pre-selected thematic aspects - namely, the relevance and degree of difficulty they attribute to the topics presented in each case, as well as the timeframes in which solutions could be ready to meet the challenge in question.

First, a summary of the discussion results on the main technical challenges for the realization of explainable AI systems is provided in section 7.1. The following section 7.2 contains a summary of the discussion results on the main regulatory challenges concerning explainable AI.

7.1 Technical challenges and need for action

Five technical challenges proved to be particularly relevant, each of which was considered both very important and solvable by almost all of the experts interviewed on the topic. The challenges are shown in Figure 10.

The topic which, according to experts, can and should be implemented in the near future is the formulation of **best practices for the selection of suitable explanatory strategies** (to which this study should also make a contribution). In the discussion it became clear that best practices are already emerging in some areas of science, especially in the field of supervised learning. There is a growing scientific literature, corresponding software prototypes and isolated success stories. However, these scientific best practices are often unusable for companies that do not maintain their own AI research departments, as the application of explainable AI is mostly related to very narrowly defined academic scenarios in these cases.

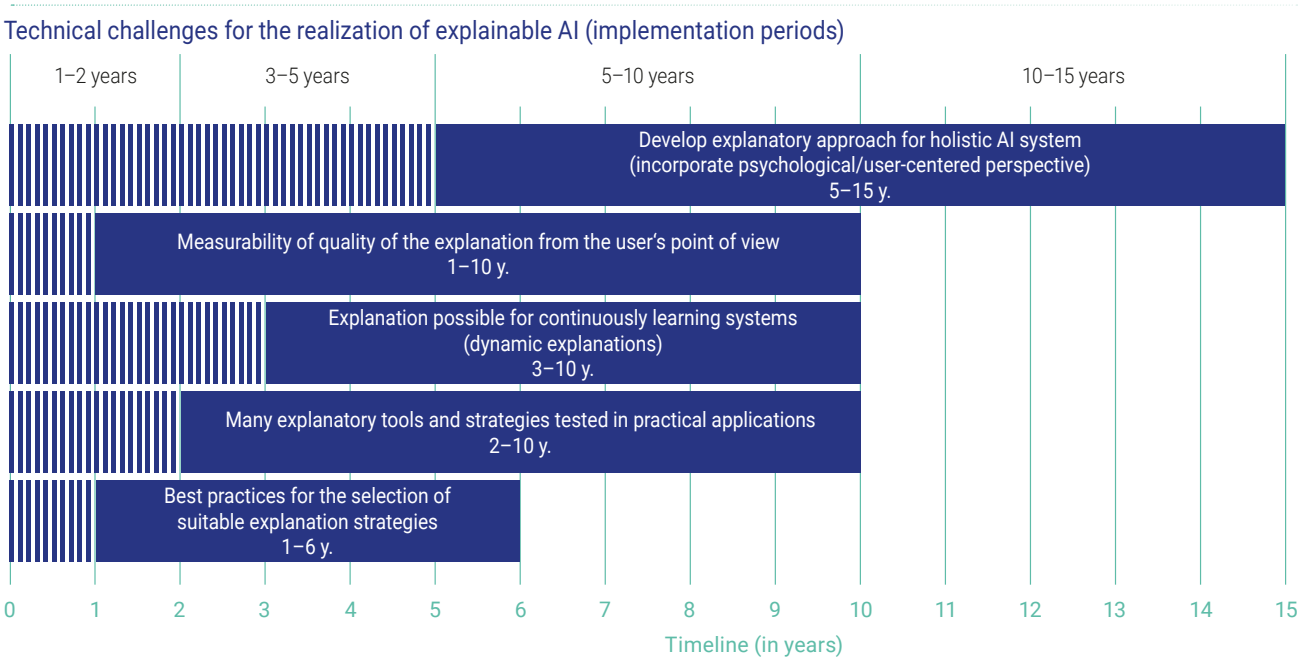


Figure 10: Result from expert interviews - Biggest challenges for the realization of explainable AI systems from a technical perspective and possible timeframes for implementation.

Typical conditions that companies face are usually not taken into account: e.g. less AI-affine users, time pressure or inhomogeneous data bases characterized by disturbances and missing data. Such factors, as well as the unclear compliance of the explanatory strategies with existing European or international regulatory and ethical frameworks, play an important role for companies, especially SMEs.

Depending on the area of application, the complexity of the problem, and the possibility of testing, the experts estimate one to five years for the development of application-related best practices. This wide range can be attributed to the very different requirements (for e.g. decision support systems or highly autonomous processes) and the lack of a unified understanding of what constitutes a "sufficient" explanation.

A large majority of experts also share the impression that many methods for explainable AI have **not yet been sufficiently tested in practice** and are therefore in many respects still the subject of research. An exception to this is the area of language-based explanations in natural language processing, where large U.S. companies such as Google and Facebook already have extensive professional and commercially successful implementations. Apart from this, however, there is often still a large gap between scientific theory and industrial implementation, especially for SMEs: For industry practitioners, the published methods are often not sufficiently well implemented and the test problems addressed in scientific articles are often not very realistic. Regardless of these obstacles, the interest of companies in explainable AI is high, according to the experts. At the same time, several experts perceive unresolved regulatory requirements as a major hurdle to the practical testing of explainable AI and AI in general. Consequently, the experts expect very different timeframes (two to ten years) for comprehensive testing of explainable AI due to the different areas of application and requirements.

Another area that was also assessed very differently by the experts due to the diversity of the AI applications and the associated requirements for explanations is the perceived **difficulty in providing explanations for continuously learning systems**. According to several experts, a technical solution is already possible today for certain applications. This is the case when simple explanatory strategies that provide specific information on individual decisions are sufficient. The same applies to pre-trained systems in which the phases of sequential (re-)training provide sufficient time to adapt the explanation strategies "offline" as well. "Dynamic" explanations that actually adapt individually to the changing system and the user, on the other hand, are considered by the experts to be much more difficult to implement. Many experts believe that solutions are only possible by expanding human-machine interaction. For systems in the latter category, several experts do not expect practical solutions for another ten years, so in general corresponding solutions are expected in three to ten years, according to the interviewees.

The majority of experts also consider the **measurability of the quality of an explanation from the user's point of view** to be an important challenge, for some it even represents a key issue. This aspect is seen as particularly relevant for establishing acceptance and comparability. In the development of suitable approaches, methods from other disciplines, such as behavioral sciences and psychology, should be taken into account. However, the concrete implementation still raises questions. An automated, algorithmic solution is seen as very difficult or a major challenge. According to the majority of experts, it is more likely that solutions will only be developed for specific applications or that a realization will require studies with users. Depending on the intended realization - studies or algorithmic – timeframes of one to ten years are expected für possible solutions.

One challenge, which includes the integration of a user-centered perspective and the previously discussed measurability of the explanatory quality, is to **develop an explanatory approach for holistic AI systems.** Some experts emphasize the consideration of approaches from psychology and cognitive sciences in this context. A major difficulty is that users should also not get a false sense of reliability, security, or safety if the AI system is not sufficiently comprehensible to them.

One of the main focuses of appropriate solutions should be to enable users to recognize at any time whether trust in individual decisions or the behavior of an AI system is justified or not. If the human blindly relies on the system's decision, there is a risk of losing problem-solving skills and technical know-how. While a majority of the experts rated the topic as important and challenging, some considered it less important. The experts expect a timeframe of 5-15 years for possible solutions.

7.2 Regulatory challenges and need for action

Various bodies at national and European level are already actively dealing with the special requirements arising from the properties of AI systems for their approval. Regulatory requirements are generally formulated in relation to applications and independently of concrete models and procedures. Nevertheless, it is obvious that the requirement for the explainability of AI systems, in particular, is largely derived from the special property of black-box AI systems - being able to make decisions without a prefabricated set of rules.

Today, in strictly regulated fields of application such as health care, the process industry, critical infrastructures, etc., there are usually no clear specifications on the part of the legislator to which the responsible approval bodies and developers could orient themselves with regard to explainability. On the other hand, if specifications do exist, they are often so challenging from a technical point of view that the use of certain AI models or methods is implicitly excluded, without this always necessarily being justified by the application. Since this has an inhibiting effect on the effective use of AI methods, regulatory challenges were discussed in the interviews with the experts with regard to their significance and feasibility. The opinion shown in Figure 11 represents a summary of the results of the discussions with all experts.

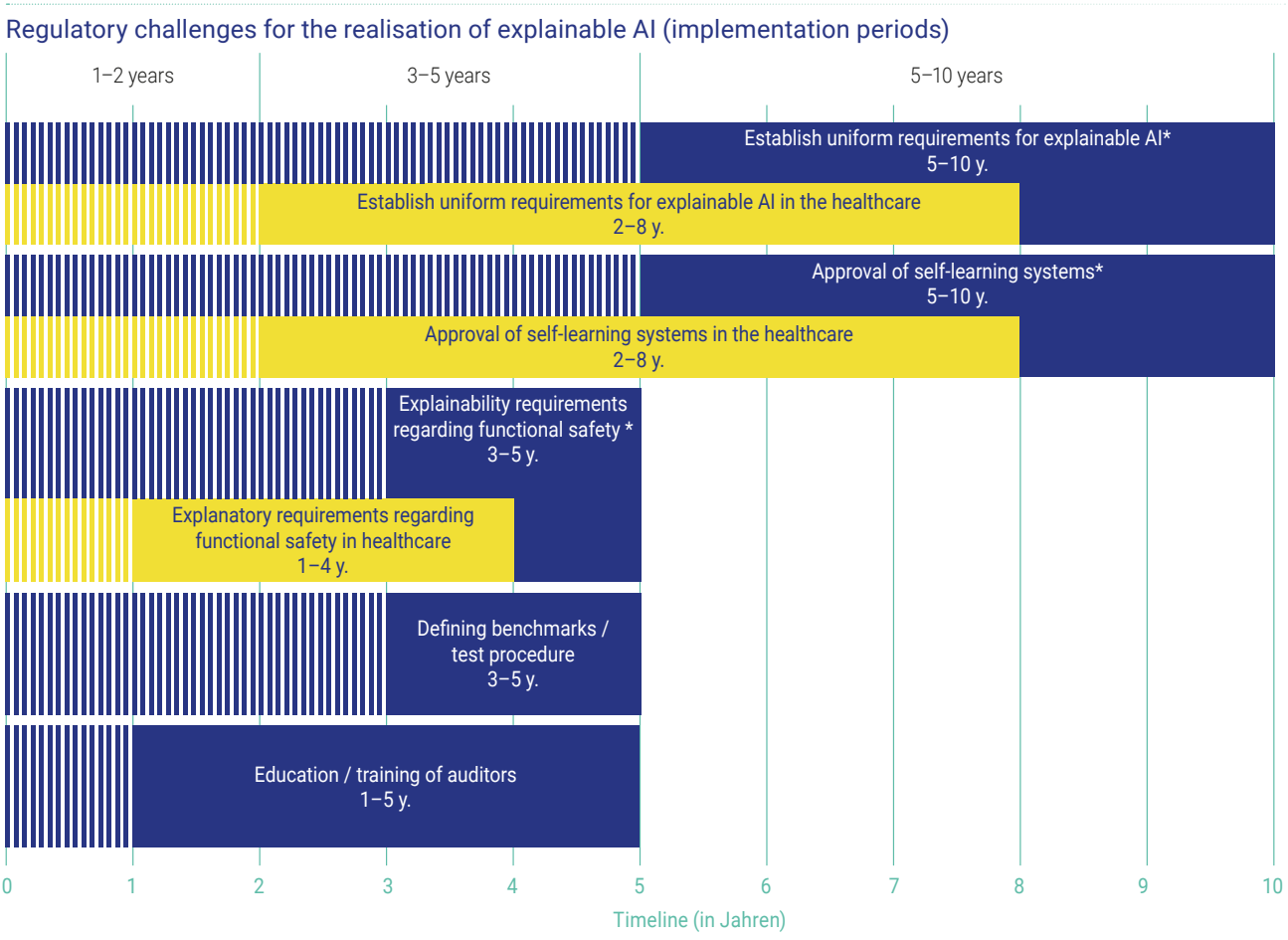


Figure 11: Result from expert interviews - greatest challenges for the realization of explainable AI systems from a regulatory/legal perspective and possible timeframes for implementation (assessment of the general development in blue and the partially faster expected development in the healthcare industry in yellow).

* The statements of the experts from the health sector are excluded here.

Five aspects proved to be particularly relevant in the discussion with the experts: Training and education of auditors, definition of benchmarks and test procedures, (explainability-related) functional safety, the definition of uniform requirements for explainable AI, and the approval of self-learning systems. There are differences in the timeframes that the experts anticipate for possible solutions to the regulatory challenges. It was noticeable in the interviews that the majority of people from the healthcare sector consider shorter cycles for the realization of individual aspects to be feasible. In the following, the general assessments and the particularities are discussed.

Among the topics, which, according to the experts, can and should be implemented in the comparatively short term, the **training and further education of the auditors** is first. Although the lack of corresponding expertise is currently perceived as a major obstacle and potential bottleneck with regard to approval procedures and although very different requirements have to be assessed depending on individual applications (including functionality, safety concepts, appropriate involvement of domain experts in decision-making processes), most experts considered implementation to be possible within three years. Those who also considered the development of suitable training programs in their domains to be time-consuming estimated three to five years.

The majority of experts considered the **establishment of benchmarks/testing procedures** a possible basis for how system and explanation behaviour could be analysed and validated by approving or certifying institutions. However, several experts pointed out that benchmarks can only be used for certain applications - e.g. autonomous driving, robotics applications - and that regular adaptation of the corresponding data sets would be absolutely necessary for such a test methodology. On the other hand, the high level of comparability and the time savings were highlighted as advantages. The experts who commented on realistic timeframes expected a period of three to five years to establish appropriate test procedures for suitable target applications.

A majority of the interviewees also considered the aspect of functional safety to be an elementary challenge for the realization of explainable AI systems from a regulatory perspective. This underscores the importance of explanations of safety aspects - regardless of whether humans are the ones who implement analysis results or proposed decisions, or corresponding actuators. In this context, one expert expressed the view that any AI system whose decisions are not reviewed sufficiently frequently by an informed user must, in principle, be considered autonomous. With regard to current regulatory frameworks, there are certain differences among the application domains and dependencies on the individual degree of automation of the application: in the process industry, for example, detailed safety concepts must be submitted in accordance with the Hazardous Incident Ordinance, irrespective of the algorithms used. Such safety concepts, which generally also include appropriate system monitoring by qualified employees, must not only be comprehensible to the approving authorities at the time of initial approval, but must also be adapted to the state of the art and recertified at regular intervals. Whereas the obligation to present and implement safety concepts in the process industry and in some cases also in manufacturing (see the Machinery Directive) is thus imposed on the plant operator, in other areas there is often no established approval practice. For the healthcare domain, where autonomous AI-based systems are not used in practice due to criticality, but only decision support systems, several experts expect clarification of functional safety requirements for explainable AI systems in one to four years.

The clear majority of experts considered setting uniform requirements for explainable AI and approving self-learning systems as key challenges for regulation.

Thereby, experts working in the healthcare industry predict that the definition of concrete requirements for explainability will occur more quickly than, for example, experts from the application areas of production and the process industry. Several experts agree that a European solution is urgently needed to **define uniform requirements for explainable AI**. Individual interviewees also emphasize the great political dimension of this challenge, as various countries and bodies in the EU must come to common agreements. The lack of uniform requirements in the affected sectors is perceived by a large amount of experts as a major obstacle to the development of explainable AI and to AI in general in Germany. This circumstance slows down approvals, the willingness to invest and the development of explainable AI in general in the affected industries. The dimension of this regulatory challenge, which also needs to involve harmonization with domain-specific regulation, is also reflected in the long predicted timeframe for a possible realization of five to ten years.

According to the experts, it is imperative to find a solution for the **approval of self-learning systems.** This raises the key question of how the relevant criteria, which are dependent on the specified requirements, can be appropriately reviewed when a system changes. This might include, among other things, checking whether a system develops ethically unacceptable decision-making strategies in the learning process, which may not be apparent in the case of initial approval. The majority of those who commented on the general matter believe that enabling responsible authorities to approve self-learning systems is a challenge, but one that can be solved. However, a lone voice felt this was not currently feasible. Several respondents saw no or only little challenge in this task, at least for clearly defined areas of application. One proposed approach included, for example, externally conducted re-certifications that could be triggered periodically after models are retrained, supported, for example, by a ticket system. However, for models that are inevitably subject to permanent change - for example, because they have to respond to changes in environmental parameters - it is difficult or impossible to define meaningful intervals for externally triggered certifications. Because quasi-continuous testing by external bodies is out of the question in such cases for a wide variety of reasons, self-certification by operating companies was proposed as a viable option.

8 CONCLUSION

8 CONCLUSION

Concepts exist to distinguish between “black-box” and “white-box” models via the three transparency levels of simulability, decomposability and algorithmic transparency. Under the assumption of comprehensible input variables, white-box models are characterized by algorithmic transparency at least. This distinguishes them decisively from black-box models, which, even in the case of low-complexity models still being of practical use, do not fulfil any of the three aforementioned tranparency properties - especially not the lowest level of algorithmic transparency. According to the literature, the decisive criterion for algorithmic transparency property is whether a model or model generation is sufficiently accessible for mathematical analysis. A classification of common procedures into the white-box and black-box categories as well as the previously mentioned transparency concepts from the literature was provided (see Chapter 2).

→ It was not surveyed in the context of the study how well known the concept of model transparency is in the respective scientific communities. The relevant literature contributions, which also refer to explainable AI, have only been published in the last five years. However, the common use of ill-defined and contradictory terms in the literature as well as scientific articles rejecting the categorization of models as black box in general or such publication doubting the suitability of opaque models (such as e. g. neural networks) for critical applications in general, indicate all one thing very clearly: The corresponding discourses are sometimes still conducted very differently in the scientific communities. A standardization of the taxonomy is still pending from the scientific side.

An examination of the established explanatory strategies and tools shows that some individual methods are designed to generate explanations only for a specific type of AI model. Others can only be used when specific types of data are used (see Chapter 3). Specific advantages and disadvantages of using each approach have been highlighted, with one thing in particular becoming apparent: If only decision explanations (local explainability) are required, established post hoc analysis tools provide opportunities to better understand black-box models, e.g. neural networks. Explanation tools such as Integrated Gradients and SHAP, which are mainly used to explain individual decisions, have already reached industrial maturity, but are not very intuitive in their handling and are therefore generally to be understood as tools for AI developers. For AI users, more intuitive approaches such as saliency maps and counterfactual explanations are often preferred.

According to the assessment of most developers and users, neural network model variants will represent the most important model type in the field of artificial intelligence in five to ten years (see Chapter 4). On average, around two thirds of those who actively use neural network variants conclude that they can already be partially explained today - at least with regard to individual decisions and when using suitable explanatory strategies. Conversely, the fact that one third of those who use or develop neural networks do not consider them to be explainable at all indicates a certain lack of awareness of existing analytical tools.

At the same time, a large majority of the persons developing or using AI systems for healthcare applications share the view that respective AI systems must be explainable if they are to be used professionally in this domain. In various other industries (finance, production, construction, process industry, energy industry, service sector), a majority of people with relevant domain knowledge also consider a certain degree of explainability to be indispensable. In these application fields, however, this is usually not due to strict approval requirements but due to potential customers and users who would simply not accept AI systems for the respective “typical” industry applications today if the AI is not explainable. The fact that, according to the survey, AI explainability will also become increasingly important for other stakeholders such as internal auditors, management and end customers in the future (see Chapter 4) underscores the perspective need for explainable AI, even beyond regulatory requirements.

→ The observations on the missing knowledge of explainability methods suggest that the scientific-methodological discourse between the disciplines of computer science (especially data science) and mathematics (especially statistics and numerical mathematics) should be fostered in basic research and education, as well as the emergence of best practices.

The comparison based on four use cases (see chapter 5) shows: Two motivational reasons for the use explainable AI are common to all four applications, namely to “find” causal relationships and to determine confidence. The first, more typical, motivation is clearly the main reason to use explainable AI in the two use cases for anomaly detection - image analysis

of histological tissue sections and machine condition monitoring. One difference between the two applications, besides the different data basis, is that in the medical case the explainability requirements are defined by the regulatory authorities, whereas the medical staff supposedly only needs decision explanations (local explainability) for operational use. In the case of machine condition monitoring, on the other hand, there are no regulatory requirements whatsoever, although users often expect model explanations (global explainability) in addition to decision explanations. The solution paths are correspondingly different: The requirement of local explainability in image analysis is addressed with the help of post hoc explanations of a black-box model, the requirement of local and global explainability with self-explanatory white-box approaches (machine condition monitoring). In the second case, however, explanations are also "actively" provided: on the one hand, in the form of statistical probabilities of occurrence (Bayesian networks) and an additional surrogate model, and on the other hand, through natural language explanations that a user can improve by herself or himself.

In the use case dealing with the text analysis of medical reports "increasing information gain" is the central motivation for using explainable AI. The additional information provided by an explainable AI system not only enables medical experts to judge whether a criterion (e.g. a symptom) that was decisive for a certain classification (e.g. similarity of disease progression of two patients) is either medically plausible or not. The "explanatory" information also enables to perform a more thorough analysis of available patient data and, possibly, to draw conclusions with regard to the medical treatment (e.g. adjustment of medication). Here, the basis for providing such decision explanations are nominal black-box models (neural networks), which are supplemented by prototypes or external knowledge bases, so that the resulting model itself can provide medically comprehensible reasons for individual decisions. In the process control use case, "determining confidences" is one of several, but ultimately the decisive overarching goal (even if intuitively it is not necessarily associated with explainability). Undetected errors in the visual state detection or susceptibility to disturbances and bias in the "hybrid" models can entail incalculable risks for the robust and stable control of the chemical plants.

Therefore, compared to the other use cases the most far-reaching explainability requirements (explainability of individual decisions and model effect mechanisms) also arise in this case. Here, the approach is to create suitable "hybrid" models from mechanistic models and simulation data as well as image and sensor data, which combine white-box and black-box components to form explainable plant models.

→ The added value of application-related case studies is clearly recognizable. The transferability of technological approaches to other ¬fields of application is comparatively easy when problems are structurally similar, e.g., in terms of data type, goals, etc. In this context, it is highly recommended to increasingly address applications that also focus on the explanation of model mechanisms (generation of global explainability through "hybrid" systems), the interaction between humans and AI systems to improve explanations ¬(such as in the use case machine condition monitoring), or explanations for (partially) autonomous systems (such as in the use case for AI-supported process control). So far, these significant application fields have only been addressed sporadically in application-oriented research.

The experts' recommendations and the findings from the literature were translated into a practical orientation guide (see Chapter 6). This is intended to provide support with regard to the first practical steps in the selection of explanatory strategies. A central finding here is that, at least for the foreseeable future, there will be a lack of explanatory tools that can provide detailed and quantitatively usable model explanations for black-box models, such as neural networks. In principle, therefore, white-box models are always to be preferred for corresponding explanatory requirements if they perform similarly well in comparison to black-box models, or at least sufficiently well with respect to the application. If model explanations and the use of black-box models are required, the use of "hybrid" approaches that combine white-box and black-box components and provide independent explanations is also promising in perspective[32]. If explanations of individual decisions are sufficient, the "orientation tree" offers a decision-making aid with regard to the explanatory strategies discussed in Chapter 3.

32 Individual procedures are currently being developed or refined in research projects and some of them were presented in the study (KEEN and Service-Meister projects of the BMWK technology programme AI Innovation Competition and the RAKI project of the BMWK technology programme Smart Data Economy).

→ Although the orientation aid takes into account the most cited approaches (in the case of their practical applicability), it thus only represents a snapshot of the state of the art. In the sense of completing best practices, it is generally advisable to continue in this direction and, in particular, to give greater consideration to quantitative comparisons and the examination of the transferability of approaches.

With regard to the technical challenges for explainable AI, it is obvious that there is yet no completed collection of best practices available that companies, especially SMEs, can take advantage of and that cover sufficient fields of application (see Chapter 7). Closely related to this is the deficit that academic scientific research often examines example applications which are far from practice and that algorithms are rarely tested on real-world problems (which is common practice e.g. in research institutions of individual high-tech companies).

→ Individual success stories and research projects (see use cases) address user requirements regarding the comprehensibility of AI already on a selective basis. However, the survey results and the statements of the experts suggest that the market demand for explainable AI will continue to grow. It is therefore recommendable to support the development of industry-specific solutions with targeted applied research activities.

→ Regarding the deficits observed by most experts and the richness of methodological approaches, it is highly recommendable to put the efficiency and user-friendliness of solutions of real problems from practice more in the center of research activities.

→ In order to strengthen the user perspective, which according to the experts is sometimes neglected, and to address the existing gap between science and industry, quantitative comparisons between alternative approaches and the state of the art should always be made whenever possible.

→ The new and further development of suitable "hybrid" approaches that combine data-driven and knowledge-driven approaches - or white-box and black-box modeling approaches - can be fostered by establishing interdisciplinary application-oriented research networks consisting of experts from computer science, mathematics (statistics, numerical analysis), and the various application disciplines.

The consideration of behavioral or cognitive science aspects of explainable AI (measurability of the quality of an explanation, explainability of holistic AI systems, automated explanation adaptations to users and self-learning systems) is seen by a majority of the interviewed experts as an important research direction and currently a major challenge.

→ The experts' statements on the user-centric topics of explainable AI and their assessment that solutions can be expected only in the medium to long term make it clear: There are various open questions here that must first be answered by basic scientific research.

In general, the technical challenges are considered surmountable by the interviewed experts. Nevertheless, as mentioned before, in many potential target industries of explainable AI, systems are subject to approval. In most of these industries, e.g. healthcare, there is a lack of clear regulatory requirements or approval guidelines to which companies can orient themselves (and align their technical developments with).

With regard to the most important regulatory challenges (see Chapter 7), the following picture emerges: The lack of uniform regulatory requirements for the explainability of AI is currently the greatest obstacle to the development of explainable AI systems. Approval processes in the affected industries are delayed and, as a result, the willingness to invest in the development of explainable AI is also slowed down. Indirectly, however, this situation also has an impact on the development of innovations in other unregulated areas, as innovations or "technology pushes" with regard to explainable AI do not occur. In the opinion of the interviewed experts, it is also clear that mechanisms for the approval and (re-)certification of self-learning systems must be found at the same time. It is a cause for concern that many experts expect regulatory requirements for explainability, which should be uniform throughout Europe, to be defined only in five to ten years. The experts with domain knowledge in healthcare expect a slightly faster progress in some areas, especially for the design of explainable AI in terms of functional safety and the physical integrity of patients. At the same time, the majority of experts, regardless of their specialization, estimate only three to five years for establishing test procedures and benchmarks.

The training and further education of examiners will be an enormously important task in the future, since these persons will have to perform many tasks of great social significance: Pre-trained systems, systems that are frequently re-trained on the basis of updated training data, and continuously learning systems must be initially approved and then regularly recertified.

This could become a bottleneck for the approval of AI products. Taking into account the fact that training programmes need to be designed, the experts estimate one to five years to implement this measure.

→ So far, there is no definition of application and risk classes from which it can be derived whether the provision of explanations is fundamentally necessary. Likewise, there is a lack of clarity in existent regulatory requirements for explainability, which should be formulated as uniformly and quantitatively as possible and refer to the application or risk class level. There will possibly be a risk classification on the basis of the proposal for the regulation of AI to be published by the EU shortly*. Nevertheless, it cannot be expected that this will provide industry- and application-specific explainability requirements or quantitative approval and certification guidelines for AI products. It is therefore recommendable – also with regard to the slow implementation processes expected by the interviewed experts – to develop specific approval and certification guidelines for AI products as soon as possible, at least for the areas of application of explainable AI that are most important from a social and economic point of view in Germany. Such a project for the development of guidelines should involve representatives from science, industry and standardization, as well as testing/certification institutions, in order to achieve the broadest possible social consensus on the one hand and to ensure practical feasibility in testing and certification on the other. ●

* At the time of the study's editorial deadline, publication was assumed to take place in April 2021, as announced by the EU Commission.

A OVERVIEW OF AI METHODS AND MODELS

A OVERVIEW OF AI METHODS AND MODELS

AI methods can be used to find or at least approximate solutions to a wide variety of classification, regression and clustering problems[33]. The terms AI and machine learning are often used synonymously. Machine learning is about an algorithm "learning" to solve a problem[34] based on training data. During the actual learning process, the degrees of freedom of a given model structure are adapted to the respective data or the specific problem. The "ssessment" of new, unknown data according to the task is then carried out via a corresponding evaluation of the adapted "AI model".

The actual fitting of the models may require the manipulation of a few or millions of parameters, depending on the specific number of degrees of freedom of the model. While the fitting of a linear regression model in a simple case only requires the determination of a single model parameter in order to relate dependent variables to independent variables, the fitting of multi-layered models that exhibit strong interconnectedness, such as neural networks, usually requires the adjustment of hundreds of parameters. If we are even dealing with so-called deep (neural) networks - i.e. if the network comprises many parameters and layers (whereby there is no precise definition for "many") - we quickly deal with millions of parameters that have to be adjusted. In this case, one also speaks of "deep learning".

Selected AI models and techniques are briefly described below..

An **autoencoder** is a neural network used for compressed encoding of data. The autoencoder consists of two components: an encoder, which compresses the input data, and a decoder, which reconstructs the original data from the compressed data. The encoder and decoder can also be used separately. Typical applications include anomaly detection or dimension reduction (Badr 2019).

Bayesian networks are probabilistic or graphical models that take the form of directed acyclic graphs whose nodes describe random variables and whose edges describe conditional probabilities. They are particularly well suited for quantifying probabilities for the possible cause of events that have occurred. Bayesian networks can be used, for example, to solve decision problems under uncertainty.

Clustering models are used to automatically divide a data set into subsets of similar data points. The commonly used K-Means-Clustering algorithm is a fast iterative algorithm that, after initially randomly selecting cluster centers, continues to adjust them so that the "clustering error" is minimized.

Neural networks frequently used in image processing are **Convolutional Neural Networks (CNNs)**. Due to the local network architecture between layers (similar to the layers in the visual cortex), their evaluation can be performed as convolution. CNNs are particularly suitable for application areas where adjacency between features plays a role, e.g. pixels in an image or words in a sentence. Similar to the visual regions of the mammalian brain, the receptive fields become larger from layer to layer and the complexity of the features to which the units respond increases.

Dimension reduction is used to reduce the number of data or features, e.g., to reduce the computational complexity of a data processing process or to extract the most important characteristics of a data set. Principal Components Analysis (PCA) is an example of a method where features are transformed unsupervised. On the other hand, the computation of the information-. content can be used to select supervised special features. Other examples include t-SNE (t-distributed stochastic neighbor embedding) and LDA (linear discriminant analysis) (Cunningham 2008; Ba-lakrishnama and Ganapathiraju

33 The goal of a classification is to assign input values to a (discrete) class or group. An example of this is image classification: images depicting animals must be assigned to one of the two classes "dog" or "cat". Regression is used to map the relationship of a dependent variable, e.g. human body size, from independent variables such as human shoe size. Clustering involves analyzing similarities and differences in data and grouping them accordingly. For example, in marketing, groups of similar products are formed (without these groups having to be known beforehand) in order to be able to present the customer with suitable offers during an online search.

34 Learning methods are often divided into supervised, unsupervised, semi-supervised and reinforcement learning methods. In supervised learning, the algorithm is given pairs of input and output values that specify what result is expected given a particular input. On this basis, the algorithm learns to correctly assign new inputs as well. In unsupervised learning, only input values are provided. The algorithm must independently recognize structures in the data. Semi-supervised learning combines both methods described above. In reinforcement learning, the algorithm independently learns a strategy based only on positive or negative feedback (Alloghani et al. 2020; Oladipupo 2010).

1998). Auto-Encoding Neural Networks can also be used to achieve dimension reduction (see above).

In **ensemble models**, rather than learning a single function or model for classification or regression based on data, several different redundant ones are learned. The results of these are then merged, for example via formation of the weighted average or majority voting. The goal of using ensemble models is to derive a "strong" classifier (strong learner) from several "weak" classifiers (base classifiers/weak learners). The individual classifiers must be both accurate and diverse. Random Forest is a well known ensemble model based on single decision trees. However, different models (e.g. decision trees and neural networks) can also be combined in an ensemble (Goos et al. 2000).

The idea behind **decision trees** is to divide a complex decision into several simple decisions. Decision trees are hierarchical structures that can be used for both classification and regression. Exemplary algorithms for creating decision trees are ID3 or C4.5. Decision trees are usually easy to understand even for laymen, since at each point within the tree it can be seen which decision has just been made (Mitchell 2010; Arrieta et al. 2019).

Expert systems (or knowledge-based systems) are knowledge databases built on the knowledge of experts (usually represented as if-then logic), which can derive conclusions and recommendations for action from the knowledge base or check the truth of statements using inference mechanisms. To build a knowledge-based system, it is necessary to have detailed knowledge of the application domain and to formalize it according to a problem-solving strategy. Expert systems are usually well understood (Karst 1992; Puppe 1988; Spreckelsen and Spitzer 2009; Wagner 2000; Lucas and Van der Gaag, Linda C. 1991).

Generative Adversarial Networks (GANs) is a neural network training method that is used especially when the amount of training data is limited. Two different networks are trained: One (called the generator) produces data that looks as realistic as possible, the other (the discriminator) tries to distinguish real data from the synthetically produced ones. Both networks are trained competitively with the goal of producing synthetic data that is as realistic as possible, i.e., data that the discriminator can no longer distinguish from the training data. These can be used to extend the training base or to complete incomplete datasets in the application (Creswell et al. 2017).

Long Short Term Memory Networks (LSTMs) belong to the group of Recurrent Neural Networks (RNNs). LSTMs are particularly well suited for processing sequences of data, e.g. speech. The advantage over RNNs is that "learned" information can be retained longer - and thus contextual information as well.

Mathematical optimization refers to a methodology for minimizing or maximizing a mathematical objective function, where variables may be subject to constraints. If objective functions and constraints are linear functions with respect to the decision variables, one speaks of linear optimization, otherwise of nonlinear optimization. The analytical solution of optimization problems is rarely possible, so that numerical solution methods are usually used to find parameters that meet the respective optimality criteria. The use of highly efficient solution methods is in fact inevitable for the solution of nonlinear optimization problems - especially if complex models are involved via constraints.

Metaheuristics can be used to explore search spaces with different strategies. Heuristics are used to find the best possible approximate solutions to optimization problems that are too complex to solve exactly. In this context, a metaheuristic is often viewed as a higher-level strategy that is used to guide "lower-level" heuristics to find suitable solutions. An example algorithm for a metaheuristic is Simulated Annealing (Bianchi et al. 2008; Voß 2001).

Neural networks consist of several layers, which in turn consist of individual units with a (usually nonlinear) transfer function that transfers the sum of the inputs into an output that is passed on to the next layer via weighted links. A network consists of an input layer, at least one hidden layer and an output layer. The number of layers is also called the depth of the network. The weights of the links and the parameters of the transfer function are adjusted as the network is trained. The complexity of neural networks - number of units and layers as well as weights of the individual connections and thus dependencies between the units - lead to the fact that these models are hardly comprehensible.

Regression models represent the relationship between one or more dependent variables and one or more independent variables. In linear regression, the assumption is made that the dependent variable is continuous and has a linear relationship with the input variables. In logistic regression, the dependent variable is considered to be binary. Logistic regression is widely used

and is also applied, for example, within neural networks (Cucchiara 2012; Karlaftis and Vlahogianni 2011). The two examples presented - linear and logistic regression models - are classified as statistical and probabilistic models, respectively. Regression models are usually easy to interpret.

The idea of **reinforcement learning is that** an "agent" interacts autonomously with its environment to achieve a goal. The autonomous agent must find its way around its environment and perform interactions for which it receives a reward, penalty, or neutral feedback from its "trainer". The agent must develop a strategy that maximizes the number of rewards. The agent has a lot of interaction possibilities and usually cannot perceive his environment completely, but only partially and possibly subject to noise. Based on the perceived environment, an action is selected and performed that changes the environment. This change is in turn perceived by the agent again. Examples of reinforcement learning algorithms include Q-learning or SARSA (Kaelbling et al. 1996; Mitchell 2010; Harmon and Harmon 1997; Sutton and Barto 2010).

Recurrent neural networks (RNNs) represent a type of neural network in which units can be connected to themselves, to units of the same layer, or to units from previous layers. This creates circles in the connectivity structure, which transforms these neural networks into dynamic systems. RNNs can represent complex dynamic relationships and have a "memory"; however, they are more computationally expensive to train for the same number of units. RNNs are used, for example, in speech recognition.

Statistical and probabilistic models have always been used to evaluate measurement data and to derive estimates and predictions about the modeled phenomena from measured and known probability distributions. The methods of statistics are to some extent similar to those of machine learning and can be used for similar goals - e.g., prediction or classification. However, such approaches require that the choice of the model is based on clear and comprehensible assumptions regarding the underlying data and processes. For data sets whose structure is very complex or unknown, "assumption-free" machine learning models are therefore often more suitable. An example for statistical models are regression models.

Support Vector Machines (SVMs) were originally developed to solve binary classification problems, but can also be used for other problems such as regression. With the help of SVMs, a nonlinear problem can be transformed into a linear one. SVMs, due to their good performance and efficient training procedures (in contrast to NNs), are often the first choice when ML problems of reasonable complexity are to be solved. However, SVMs themselves can also become very complex, such that the lack of comprehensibility can be a concern.

Transformer Networks are a type of neural networks that are specifically applied in language processing, for example for translation tasks, the generation of texts or summaries. Transformer Networks consist of two components: the encoder and the decoder. The encoder creates a representation of an input, then the decoder generates a corresponding output word by word (Uszkoreit 2017).

The basis for **knowledge graphs and Semantic Web technologies** are models for representing knowledge that can be read and "understood" by computers. Purely data-driven ML methods have the problem that they can only recognize patterns in the data, but cannot put these patterns into the broader real-world context. Knowledge graphs can provide this context. These models are used as the basis for reusing the "linked" information in intelligent systems. The goal of the Semantic Web is to identify more detailed information more quickly and to increase semantic interoperability on the Internet. This is done with the help of ontologies, which in turn specify concepts that are essential or important within an application domain. Different languages are developed to represent these ontologies: Examples are the Resource Description Framework (RDF), the Web Ontology Language (OWL) and the Rule Interchange Format (RIF) (Hitzler 2008).

BIBLIOGRAPHY

BIBLIOGRAPHY

acatech (2020): Machine Learning in der Medizintechnik. Analyse und Handlungsempfehlungen. Munich: acatech (acatech Position). Available online at https:// www.acatech.de/publikation/machine-learning-in-der-medizintechnik/, last checked 15.03.2022.

Adadi, Amina; Berrada, Mohammed (2018): Peeking Inside the Black Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access (Volume: 6). Available online at https://ieeexplore.ieee.org/document/8466590, last checked 15.03.2022.

Alloghani, Mohamed; AlJumeily, Dhiya; Mustafina, Jamila; Hussain, Abir; Aljaaf, Ahmed J. (2020): A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In: Michael W. Berry, Azlinah Mohamed and Bee Wah Yap (eds.): Supervised and Unsupervised Learning for Data Science, vol. 9. Cham: Springer International Publishing (Unsupervised and Semi-Supervised Learning), pp. 3-21. Available online at https://link.springer.com/chapter/10.1007%2F978-3-030-22475-2_1, last checked 15.03.2022.

Working Group Health Care, Medical Technology, Care (2019): Learning Systems in the Healthcare System. Hg. v. Plattform Lernende Systeme, Deutsche Akademie der Wissenschaften. Munich. English language version of Executive Summary available online at https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen_EN/AG6_Executive_Summary_final_200206.pdf, last checked 15.03.2022.

Arrieta, Alejandro Barredo; Diaz-Rodriguez, Natalia; Ser, Javier Del; Bennetot, Adrien; Tabik, Siham; Barbado, Alberto et al. (2019): Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI. Available online at http://arxiv.org/pdf/1910.10045v2, last checked 15.03.2022.

Bach, Sebastian; Binder, Alexander; Montavon, Grégoire; Klauschen, Frederick; Müller, Klaus-Robert; Samek, Wojciech: On PixelWise Explanations for NonLinear Classifier Decisions by LayerWise Relevance Propagation. In: PloS One (7). Available online at https://pubmed.ncbi.nlm.nih.gov/26161953/, last checked 15.03.2022.

Badr, Will (2019): Auto-Encoder: What Is It? And What Is It Used For? (Part 1). Available online at https://towardsdatascience.com/auto-encoder-what-is-it-and-what-is-it-used-for-part-1-3e5c6f017726, last checked 15.03.2022.

Baehrens, David; Schroeter, Timon; Harmeling, Stefan; Kawanabe, Motoaki; Hansen, Katja; Mueller, Klaus-Robert (2009): How to Explain Individual Classification Decisions. Available online at http://arxiv.org/pdf/0912.1128v1, last checked 15.03.2022.

Balakrishnama, S.; Ganapathiraju, Aravind (1998): Linear Discriminant Analysis A Brief Tutorial. Available online at https://www.researchgate.net/publication/240093048_Linear_Discriminant_Analysis-A_Brief_Tutorial, last checked 15.03.2022.

Barbalau, Antonio; Cosma, Adrian; Ionescu, Radu Tudor; Popescu, Marius (2020): A Generic and Model-Agnostic Exemplar Synthetization Framework for Explainable AI. Available online at http://arxiv.org/pdf/2006.03896v3, last checked 15.03.2022.

BDVA Task Force 7 - Sub-group Healthcare (2020): AI In Healthcare Whitepaper. Ed. v. Big Data Value Association. Available online at https://www.bdva. eu/sites/default/files/AI%20in%20Healthcare%20Whi-tepaper_November%202020_0.pdf, last checked 15.03.2022.

Bhatt, Umang; Xiang, Alice; Sharma, Shubham; Weller, Adrian; Taly, Ankur; Jia, Yunhan et al. (2019): Explainable Machine Learning in Deployment. Available online at http://arxiv.org/pdf/1909.06342v4, last checked 15.03.2022.

Bianchi, Leonora; Dorigo, Marco; Gambardella, Luca Maria; Gutjahr, Walter J. (2008): A survey on metaheuristics for stochastic combinatorial optimization. In: Natural Computing volume 8, pages 239-287. Available online at https://link.springer.com/article/10.1007%2Fs11047-008-9098-4, last checked 15.03.2022.

Bundesinstitut für Arzneimittel und Medizinprodukte (o. D.): Benannte Stelle. Available online at https://www.dimdi.de/dynamic/de/medizinprodukte/institutionen/benannte-stellen/, last checked on 15.03.2022.

Chattopadhyay, Aditya; Sarkar, Anirban; Howlader, Pran-tik; Balasubramanian, Vineeth N. (2017): Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. Available online at http://arxiv.org/pdf/1710.11063v3, last checked 15.03.2022.

Cortez, Paulo; Embrechts, Mark J. (2011): Opening black box data mining models using sensitivity analysis. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). Available online at https://ieeexplore.ieee.org/document/5949423, last checked 15.03.2022.

Creswell, Antonia; White, Tom; Dumoulin, Vincent; Arulkumaran, Kai; Sengupta, Biswa; Bharath, Anil A. (2017) Generative adversarial networks: an overview. In: IEEE Signal Process. Mag. (IEEE Signal Processing Magazine) (1). Available online at http://arxiv.org/pdf/1710.07035v1, last checked 15.03.2022.

Cucchiara, Andrew (2012): Applied Logistic Regression. Available online at https://www.researchgate.net/publication/261659875_Applied_Logistic_Regression, last checked 15.03.2022.

Cunningham, Pádraig (2008): Dimension Reduction. In: Matthieu Cord and Pádraig Cunningham (eds.): Machine Learning Techniques for Multimedia, vol. 12. Berlin, Heidelberg: Springer Berlin Heidelberg (Cognitive Technologies), pp. 91-112. Available online at https://link.springer.com/chapter/10.1007%2F978-3-540-75171-7_4, last checked 15.03.2022.

Danilevsky, Marina; Qian, Kun; Aharonov, Ranit; Katsis, Yannis; Kawas, Ban; Sen, Prithviraj (2020): A Survey of the State of Explainable AI for Natural Language Processing. Available online at http://arxiv.org/pdf/2010.00711v1, last checked 15.03.2022.

Datta, Anupam; Sen, Shayak; Zick, Yair (2016): Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems (IEEE Symposium on Security and Privacy 2016). Available online at https://ieeexplore.ieee.org/document/7546525, last checked 15.03.2022.

Doshi-Velez, Finale; Kim, Been (2017) Towards A Rigorous Science of Interpretable Machine Learning. Available online at http://arxiv.org/pdf/1702.08608v2, last checked 15.03.2022.

eco - Verband der Internetwirtschaft e.V.. (ed.) (2019): Künstliche Intelligenz. Intelligenz. Potenzial und nachhaltige Veränderung der Wirtschaft in Deutschland. Available online at https://www.eco.de/kuenstliche-intelligenz-potenzial-und-nachhaltige-veraenderung-der-wirtschaft-in-deutschland/#download, last checked on 15.03.2022.

Erhan, Dumitru; Bengio, Y.; Courville, Aaron; Vincent, Pascal (2009): Visualizing Higher-Layer Features of a Deep Network. Université de Montréal. Available online at https://www.researchgate.net/profile/Aaron_Courville/publication/265022827_Visualizing_Higher-Layer_Features_of_a_Deep_Network/links/53ff82b00cf24c81027da530.pdf, last checked 15.03.2022.

European Commission (ed.) (2018): Commission Staff Working Document - Evaluation of the Machinery Directive. Brussels. Available online at https://ec.europa.eu/transparency/regdoc/rep/10102/2018/EN/SWD-2018-160-F1-EN-MAIN-PART-1.PDF, last checked 15.03.2022.

European Commission (ed.) (2020): White Paper: On Artificial Intelligence. A European approach to excellence and trust. Available online at https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, last checked 15.03.2022.

Gilpin, Leilani H.; Bau, David; Yuan, Ben Z.; Bajwa, Ayesha; Specter, Michael; Kagal, Lalana (2018) Explaining Explanations: An Overview of Interpretability of Machine Learning. Available online at http://arxiv.org/pdf/1806.00069v3, last checked 15.03.2022.

Gondal, Waleed M.; Köhler, Jan M.; Grzeszick, René; Fink, Gernot A.; Hirsch, Michael (2017): Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. Available online at http://arxiv.org/pdf/1706.09634v1, last checked 15.03.2022.

Google (ed.) (2020): AI Explanations Whitepaper. Available online at https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf, last checked 15.03.2022.

Goos, Gerhard; Hartmanis, Juris; Leeuwen, Jan (2000): Multiple Classifier Systems. First International Workshop, MCS 2000 Cagliari, Italy, June 21-23, 2000 Proceedings. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science, 1857). Available online at http://dx.doi.org/10.1007/3-540-45014-9, last checked 15.03.2022.

Goyal, Yash; Wu, Ziyan; Ernst, Jan; Batra, Dhruv; Parikh, Devi; Lee, Stefan (2019): Counterfactual Visual Explanations. Available online at http://arxiv.org/pdf/1904.07451v2, last checked 15.03.2022.

Harmon, Mance E.; Harmon, Stephanie S. (1997): Reinforcement Learning: A Tutorial. Available online at https://www.semanticscholar.org/paper/Reinforcement-Learning%3A-A-Tutorial.-Harmon-Harmon/5845d0a23fd5ed96c03428d6da4e99b32de8b3f1, last checked 15.03.2022.

High-Level Expert Group on AI (ed.) (2019): Ethics guidelines for trustworthy AI. Available online at https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1, last checked 15.03.2022.

Hitzler, Pascal (2008): Semantic Web. Grundlagen. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (eXamen.press). Available online at http://dx.doi.org/10.1007/978-3-540-33994-6, last checked 14.04.2021.

Holzinger, Andreas (2018): Explainable AI (ex-AI). Ed. by Informatik Spektrum. Available online at https://www.springerprofessional.de/explainable-ai-ex-ai/15586620, last checked 19/02/2021.

Holzinger, Andreas; Biemann, Chris; Pattichis, Constantinos S.; Kell, Douglas B. (2017): What do we need to build explainable AI systems for the medical domain? Available online at http://arxiv.org/pdf/1712.09923v1, last checked 14.04.2021.

Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland (ed.) (2020): Fragenkatalog "Künstliche Intelligenz bei Medizinprodukten". Available online at http://www.ig-nb.de/dok_view?oid=795601, last checked on 15.03.2022.

Kaelbling, L. P.; Littman, M. L.; Moore, A. W. (1996): Reinforcement Learning: A Survey. In: jair (Journal of Artificial Intelligence Research). Available online at https://www.jair.org/index.php/jair/article/view/10166, last checked 15.03.2022.

Karlaftis, M. G.; Vlahogianni, E. I. (2011): Statistical methods versus neural networks in transportation research: differences, similarities and some insights. In: Transportation Research Part C: Emerging Technologies (3). Available online at https://www.sciencedirect.com/science/article/abs/pii/S0968090X10001610?via%3Di-hub, last checked 15.03.2021.

Karst, Michael (1992): Methodische Entwicklung von Expertensystemen. Wiesbaden, s.l.: Deutscher Universitätsverlag (DUV Wirtschaftswissenschaft). Available online at http://dx.doi.org/10.1007/978-3-663-14584-4, last checked 14.04.2021.

Kawaguchi, Kenji (2016): Deep Learning without Poor Local Minima. Available online at http://arxiv.org/pdf/1605.07110v3, last checked 14.04.2021.

Li, Oscar; Liu, Hao; Chen, Chaofan; Rudin, Cynthia (2017): Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. Available online at http://arxiv.org/pdf/1710.04806v2, last checked 15.03.2022.

Lipton, Zachary C. (2016): The Myth of Model Interpretability. Available online at http://arxiv.org/pdf/1606.03490v3, last checked 14.04.2021.

Lucas, Peter J.; Van der Gaag, Linda C. (1991): Principles of expert systems. Wokingham: AddisonWesley. Available online at https://www.researchgate.net/publication/224818110_Principles_of_Expert_Systems, last checked 15.03.2022.

Lundberg, Scott; Lee, SuIn (2017): A Unified Approach to Interpreting Model Predictions. Available online at http://arxiv.org/pdf/1705.07874v2, last checked 15.03.2022.

Mangalathu, Sujith; Hwang, Seong-Hoon; Jeon, Jong-Su (2020): Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Engineering Structures. Available online at https://www.sciencedirect.com/science/article/abs/pii/S0141029620307513?via%3Di-hub, last checked 15.03.2022.

Mazzanti, Samuele (2020): SHAP values explained exactly how you wished someone explained to you. Available online at https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30, last checked 15.03.2022.

Mitchell, Tom M. (2010): Machine learning. International ed., [Reprint.] New York, NY: McGraw-Hill (McGraw-Hill series in computer science).

Molnar, Christoph (2019): Interpretable machine learning. A guide for making black box models explainable. 1st edition. Available online at https://christophm.github.io/interpretable-ml-book/, last checked 15.03.2022.

Montavon, Grégoire; Binder, Alexander; Lapuschkin, Sebastian; Samek, Wojciech; Müller, Klaus-Robert (2019): Layer-Wise Relevance Propagation: An Overview. In: Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen and Klaus-Robert Müller (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer International Publishing, pp. 193-209. Available online at https://doi.org/10.1007/978-3-030-28954-6_10, last checked 15.03.2022.

Nagpal, Kunal; Foote, Davis; Liu, Yun; Po-Hsuan; Chen; Wulczyn, Ellery et al. (2018): Development and Validation of a Deep Learning Algorithm for Improving Gleason Scoring of Prostate Cancer. In: npj Digit. Med (1). Available on-line at http://arxiv.org/pdf/1811.06497v1, last checked 15.03.2022.

Nambiar, Ananthan; Liu, Simon; Hopkins, Mark; Heflin, Maeve; Maslov, Sergei; Ritz, Anna (2020): Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks (13). Available online at https://www.biorxiv.org/content/10.1101/2020.06.15.153643v1, last checked 15.03.2022.

Nguyen, Daria (2020): Explain Your ML Model Predictions With Local Interpretable Model-Agnostic Explanations (LIME). Ed. v. Publicis Sapient Engineering. Available online at https://medium.com/xebia-france/explain-your-ml-model-predictions-with-local-interpretable-model-agnostic-explanations-lime-82343c5689db, last checked 15.03.2022.

Oladipupo, Taiwo (2010): Types of Machine Learning Algorithms. In: Yagang Zhang (Ed.): New Advances in Machine Learning: InTech. Available online at https://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms, last checked 15.03.2022.

Otter, Daniel W.; Medina, Julian R.; Kalita, Jugal K. (2018): A Survey of the Usages of Deep Learning in Natural Language Processing. Available online at http://arxiv.org/pdf/1807.10854v3, last checked 15.03.2022.

Pocevičiūtė, Milda; Eilertsen, Gabriel; Lundström, Claes (2020): Survey of XAI in digital pathology. Available online at http://arxiv.org/pdf/2008.06353v1, last checked 15.03.2022.

Puppe, Frank (1988): Einführung in Expertensysteme. Berlin, Heidelberg: Springer (Studienreihe Informatik). Available online at http://dx.doi.org/10.1007/978-3-662-00706-8, last checked 15.03.2022.

Ribeiro, Marco Tulio; Singh, Sameer; Guestrin, Carlos (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Available online at http://arxiv.org/pdf/1602.04938v3, last checked 15.03.2022.

Rudin, Cynthia (2019): Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. In: Nat Mach Intell. Available online at http://arxiv.org/pdf/1811.10154v3, last checked 15.03.2022.

Salehi, Mohammadreza (2020): A Review of Different Interpretation Methods in Deep Learning (Part 2: Pixel-wise Decomposition, DeepLIFT, LIME). Available online at https://medium.com/@mrsalehi/a-review-of-different-interpretation-methods-in-deep-learning-part-2-in-put-gradient-layerwise-e077609b6377, last checked 15.03.2022.

Samek, Wojciech; Montavon, Grégoire; Vedaldi, Andrea (2019): Explainable AI. Interpreting, explaining and visualizing deep learning (Lecture notes in computer series Lecture notes in artificial intelligence). Available online at https://link.springer.com/book/10.1007/978-3-030-28954-6, last checked 15.03.2022.

Schaaf, Nina; Wiedenroth, Saskia Johanna; Wagner, Philipp (2021): Erklärbare KI in der Praxis: Anwendungsorientierte Evaluation von xAI-Verfahren. Edited by Marco Huber and Werner Kraus. Available online at https://www.ki-fortschrittszentrum.de/de/studien/erklaerbare-ki-in-der-praxis.html, last checked 15.03.2022.

Selvaraju, Ramprasaath R.; Cogswell, Michael; Das, Abhishek; Vedantam, Ramakrishna; Parikh, Devi; Batra, Dhruv (2019): Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: Int J Comput Vis (International Journal of Computer Vision) (2). Available online at http://arxiv.org/pdf/1610.02391v4, last checked 15.03.2022.

Shiebler, Dan (2017): Understanding Neural Networks with Layerwise Relevance Propagation and Deep Taylor Series. Available online at http://danshiebler.com/2017-04-16-deep-taylor-lrp/, last checked 15.03.2022.

Shrikumar, Avanti; Greenside, Peyton; Kundaje, Anshul (2017): Learning Important Features Through Propagating Activation Differences. In: PMLR 70:31453153. Available on-line at http://arxiv.org/pdf/1704.02685v2, last checked 15.03.2022.

Shrikumar, Avanti; Greenside, Peyton; Shcherbina, Anna; Kundaje, Anshul (2016): Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. Available online at http://arxiv.org/pdf/1605.01713v3, last checked 15.03.2022.

Sokol, Kacper; Flach, Peter (2019): Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. Conference on Fairness, Accountability, and Transparency (FAT* '20), January 2730, 2020, Barcelona, Spain. Available online at http://arxiv.org/pdf/1912.05100v1, last checked 15.03.2022.

Spreckelsen, Cord; Spitzer, Klaus (2009): Wissensbasen und Expertensysteme in der Medizin. KI-Ansätze zwischen klinischer Entscheidungsunterstützung und medizinischem Wissensmanagement. 1st ed. Wiesbaden: Vieweg+Teubner Verlag / GWV Fachverlage GmbH Wiesbaden (Medical Informatics). Available online at http://dx.doi.org/10.1007/978-3-8348-9294-2, last checked 15.03.2022.

Springenberg, Jost Tobias; Dosovitskiy, Alexey; Brox, Thomas; Riedmiller, Martin (2014): Striving for Simplicity: The All Convolutional Net. Available online at http://arxiv.org/pdf/1412.6806v3, last checked 15.03.2022.

Stepin, Ilia; Alonso, Jose M.; Catala, Alejandro; Pereira-Farina, Martin (2021): A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. In: IEEE Access. Available on-line at https://ieeexplore.ieee.org/document/9321372, last checked 15.03.2022.

Strong Medicine (2018): An Approach to Chest Pain, Jan. 29, 2018. Available online at https://www.youtube.com/watch?v=-i67erljNYI, last checked 15.03.2022.

Sundararajan, Mukund; Taly, Ankur; Yan, Qiqi (2017): Axiomatic Attribution for Deep Networks. Available online at http://arxiv.org/pdf/1703.01365v2, last checked 15.03.2022.

Sutton, Richard S.; Barto, Andrew G. (2010): Reinforcement learning. An introduction. [Reprint]. Cambridge, Mass.: MIT Press (A Bradford book). Available online at https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf, last checked 15.03.2022.

Tjoa, Erico; Guan, Cuntai (2020): A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. In: IEEE Trans. Neural Netw. Learning Syst. (IEEE Transactions on Neural Networks and Learning Systems). Available online at http://arxiv.org/pdf/1907.07374v5, last checked 15.03.2022.

Touretzky, David S. (ed.) (1996): Advances in neural information processing systems 8. Proceedings of the 1995 conference ; [papers presented at the Ninth Annual Conference on Neural Information Processing Systems (NIPS), held in Denver, Colorado from Nov. 27 to Nov. 30, 1995. Conference on Neural Information Processing Systems; Annual Conference on Neural Information Processing Systems; NIPS. Cambridge, Mass.: MIT Press. Available online at https://mitpress.mit.edu/books/advances-neural-information-processing-systems-8, last checked 15.03.2022.

U.S. Food & Drug Administration (Ed.) (2020): Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). Discussion Paper and Request for Feedback. Available online at https://www.fda.gov/media/122535/download, last checked 15.03.2022.

U.S. Food & Drug Administration (2021): Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. Available online at https://www.fda.gov/media/145022/download, last checked 15.03.2022.

Uszkoreit, Jakob (2017): Transformer: A Novel Neural Network Architecture for Language Understanding. Available on-line at https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html, last checked 15.03.2022.

van Aken, Betty; Papaioannou, Jens-Michalis; Mayrdorfer, Manuel; Budde, Klemens; Gers, Felix A.; Löser, Alexander (2021): Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. Available online at https://arxiv.org/pdf/2102.04110v1.pdf, last checked 15.03.2022.

VERBAND DER CHEMISCHEN INDUSTRIE e.V.. (Ed.) (2012): Leitfaden zur Anwendung der Maschinenverordnung in Anlagen der chemisch-pharmazeutischen Industrie. Available online at https://www.vci.de/langfassungen/langfassungen-pdf/210608-vci-leitfaden-maschrl.pdf, last checked on 15.03.2022.

Voß, Stefan (2001): Meta-heuristics: The State of the Art. In: G. Goos, J. Hartmanis, J. van Leeuwen and Alexander Nareyek (eds.): Local Search for Planning and Scheduling, vol. 2148. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture Notes in Computer Science), pp. 1-23. Available online at https://link.springer.com/chapter/10.1007%2F3-540-45612-0_1, last checked 15.03.2022.

Wachter, Sandra; Mittelstadt, Brent; Russell, Chris (2017): Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. In: Harvard Journal of Law & Technology. Available online at http://arxiv.org/pdf/1711.00399v3, last checked 15.03.2022.

Wagner, Marc (2000): Bayes-Netze. An introduction. Available on-line at https://itp.uni-frankfurt.de/~mwagner/talks/Bayes.pdf, last checked 15.03.2022.

Wahlster, Wolfgang; Winterhalter, Christoph (eds.) (2020): Deutsche Normungsroadmap Künstliche Intelligenz. DIN/DKE. Available online at https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf, last checked 15.03.2022.

Wolf, Thomas; Debut, Lysandre; Sanh, Victor; Chaumond, Julien; Delangue, Clement; Moi, Anthony et al. (2019): HuggingFace's Transformers: State-of-the-art Natural Language Processing. Available online at https://arxiv.org/pdf/1910.03771v5.pdf, last checked 15.03.2022.

Ye, Andre (2020): Every ML Engineer Needs to Know Neural Network Interpretability. Available online at https://towardsdatascience.com/every-ml-engineer-needs-to-know-neural-network-interpretability-afea2ac0824e, last checked 15.03.2022.

Zeiler, Matthew D.; Fergus, Rob (2013): Visualizing and Understanding Convolutional Networks. Available online at https://arxiv.org/pdf/1311.2901v3.pdf, last checked 15.03.2022.

Zeiler, Matthew D.; Taylor, Graham W.; Fergus, Rob (2011): Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 International Conference on Computer 06.11.2011 13.11.2011. Available online at https://ieeexplore.ieee.org/document/6126474, last checked 15.03.2022.

Zhou, Bolei; Khosla, Aditya; Lapedriza, Agata; Oliva, Aude; Torralba, Antonio (2015): Learning Deep Features for Discriminative Localization. Available online at https://arxiv.org/pdf/1512.04150v1.pdf, last checked on 15.03.2022.