Lernende Systeme

GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

Explainable AI

White Paper

Samek, W., Schmid, U. et al.
WG Technological Enablers
and Data Science

Executive Summary

Artificial intelligence (AI) is used in many areas of society such as medical diagnostics, financial decision making, or automated quality control in manufacturing. With its increasing prevalence, the demand grows for making the functioning and decision-making processes of AI systems transparent and explainable. After all, how and why ChatGPT and other AI-based systems arrive at their results, why an instruction to a chatbot leads to exactly the sequence of words it outputs, often remains opaque. For many fields of application, particularly in safety-critical or ethically sensitive areas such as medicine or finance, transparency and explainability are crucial to contextualize and critically assess the results.

In general, highly complex AI systems are often so-called black boxes. Due to the intricacy of the models, it is often no longer possible to trace the internal functioning of the model or what exactly happens between input and output, especially in the field of machine learning. This lack of transparency can not only undermine users' trust but also hinder the further development and improvement of AI systems. This is precisely where the concept of Explainable AI (short: XAI) comes into play.

Initial Situation

In recent years, research in the field of XAI has made significant progress and has developed into a promising area of study. XAI aims to improve the safety, quality, and trustworthiness of AI systems by, for example, uncovering biases or enabling concrete model improvements. In this way, XAI contributes to fostering trust as well as values such as transparency and reliability in AI systems. These are values that are also promoted in regulatory frameworks such as the General Data Protection Regulation (GDPR) or the AI Act, as well as in standards of the International Organization for Standardization (ISO) and in the evaluation criteria of the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS). At the same time, the use of XAI offers companies the opportunity to increase their productivity and thus their profits by, for instance, detecting errors in AI applications more quickly, optimizing these applications, or strengthening consumer trust.

Explainability should always be embedded in user interfaces and specific contexts to ensure that AI systems can serve human needs in a responsible manner. Not all users require the same type of explanation. Developers, for example, need technical information to improve AI systems, while consumers are more interested in the reasons for a specific decision, such as the approval of a loan. XAI should therefore follow a human-centered approach in which comprehensibility is tailored to the respective needs, knowledge, and responsibilities of the target group, from AI specialists to everyday users.

Understanding of Concepts in the XAI Context

What exactly is meant by transparency, and how does this concept differ from explainability? A clear definition and distinction of central concepts related to XAI is necessary, since no universally accepted taxonomy exists to date. **Transparency**, understood in a rather narrow technical sense, refers to a system property that makes the inner workings of the system, including its functioning, internal inferences, and processing steps, comprehensible and also adequately explains system inputs and outputs, such as system decisions. **Explainability** is about providing users with reasons for the behavior of an AI model so that they can understand its outcomes, for instance predictions or classifications. Since this XAI-method often lies outside the model itself, it does not explain the internal processes of the model. **Interpretability**, on the other hand, refers directly to the model itself, often called a white-box model. Such a model inherently follows a decision-making process that is understandable to humans and thus provides its own explanation for a classification or prediction. **Comprehensibility**, as an overarching goal, is achieved through both interpretability and explainability.

XAI Along the Question: What Should Be Explained, to Whom, for What Purpose, and How?

How can the black box be opened with the help of XAI? The key lies in a reflective use of XAI guided by the question: What Should Be Explained, to Whom, for What Purpose, and How? XAI is therefore not an end in itself and does not provide "one-size-fits-all solutions". Rather, it is important to develop and apply XAI with realistic and appropriate expectations, since it can ultimately explain only certain aspects of model behavior. It cannot provide a complete and exhaustive explanation, which in many cases is not necessarily required.

Object of Explanation – What Should Be Explained? Initially, the question arises as to what exactly should be explained (regarding a certain target group, see below). In the process of machine learning, the object of explanation can lie on the input level, the model level, or the output level. XAI-methods provide valuable insights at each of these three levels, whether concerning data quality, model understanding, or the explanation of specific classifications and predictions. At the input level, faulty or biased datasets can be uncovered. At the model level, for example in deep learning, it is possible to analyze which concepts are represented by individual neurons in an artificial neural network. At the output level, one can determine how a specific model output was generated, such as the response of a chatbot to a query or a suggestion for a medical diagnosis. The three levels can either serve as separate elements of an explanation or, with the help of more recent methods such as SemanticLens, be combined into a holistic approach in order to systematically assess and validate the role and functioning of all AI components.

Target Group – To Whom Should It Be Explained? The question of whom refers to the specific target group of an explanation. Explanations must be adapted to this audience, since expectations and requirements vary greatly across different groups. Factors such as prior knowledge in the field of AI, time pressure, motivation, and interests all play a role. For model creators, for example, metrics regarding the fidelity of explanations may be important, while for end users the comprehensibility and usefulness of an explanation are more relevant, with these criteria being heavily context-dependent.

In general, three target groups can be distinguished. The first group, model developers and AI researchers, requires technical knowledge and detailed explanations to improve and validate AI models using XAI methods. The second group, AI laypersons and end users, values trustworthiness and expects simple and accessible explanations. The third group, such as product developers, certifiers, and domain researchers, expects XAI methods to support them in their specific activities. For this group, trustworthiness and quality improvement of AI are of central importance.

Goals of Explanation – For What Purpose Should It Be Explained? This question clarifies the concrete goals of explanation. These goals should be realistic to determine when XAI is meaningful and when it is not. The use of XAI methods and approaches primarily reveals two broader goals of XAI, which may partially overlap. On the one hand, XAI can be used to meet the dimensions of trustworthy AI such as control, fairness, and reliability. The aim here is to strengthen trust in AI. On the other hand, XAI can be employed as an engineering tool to improve the quality of AI models, for instance by enhancing the data basis, refining the models, or enabling domain-specific adaptation. Although the different target groups of XAI each have their own reasons for using it, for example to learn something about the data and a specific problem, to develop new antibiotics, or to make an important financial or existential decision more understandable, the two goals of XAI for trustworthy AI and XAI as an engineering tool largely serve as means to realize these target-group-specific objectives.

Implementation of Explanation – How Should It Be Explained? The how-question concerns the form of the explanation, the choice of suitable XAI methods and the correctness and usefulness. This helps to better understand and optimize the effectiveness of an AI model in practical use.

The **form** of an explanation must be tailored to the target group, understandable, and context-specific, whether visual, verbal, or multimodal. Examples include visual highlighting critical decisions in time or detailed verbal explanations in complex situations. Furthermore, the XAI **methods** describe how it is technically generated, for instance through relevance-based, concept-based, or example-based methods. Newer approaches such as interactive XAI actively involve users in the machine learning and explanation process, and generative AI poses new challenges for XAI due to the sheer size of models and datasets. Mechanistic interpretability also belongs to these newer approaches, as it seeks to systematically analyze and interpret the internal structures of neural networks.

Since explanations are not always correct, **evaluation** using specific metrics is often necessary, making it possible to systematically assess and improve XAI explanations. Among these are measures of fidelity and robustness, meaning the alignment of the explanation with the actual model behavior on the one hand and the stability of explanations when models are subjected to small input perturbations on the other.

Illustration of relevance-based, example-based, or concept-based methods

Relevance-based

The XAI component of the AI system gives users an explanation of the output based on the pixels relevant for the classification. Users recognise that some of the pixels are irrelevant here.

Example-based

The XAI component of the AI system gives users a reference example. The user recognises that the reference example is incorrectly classified.

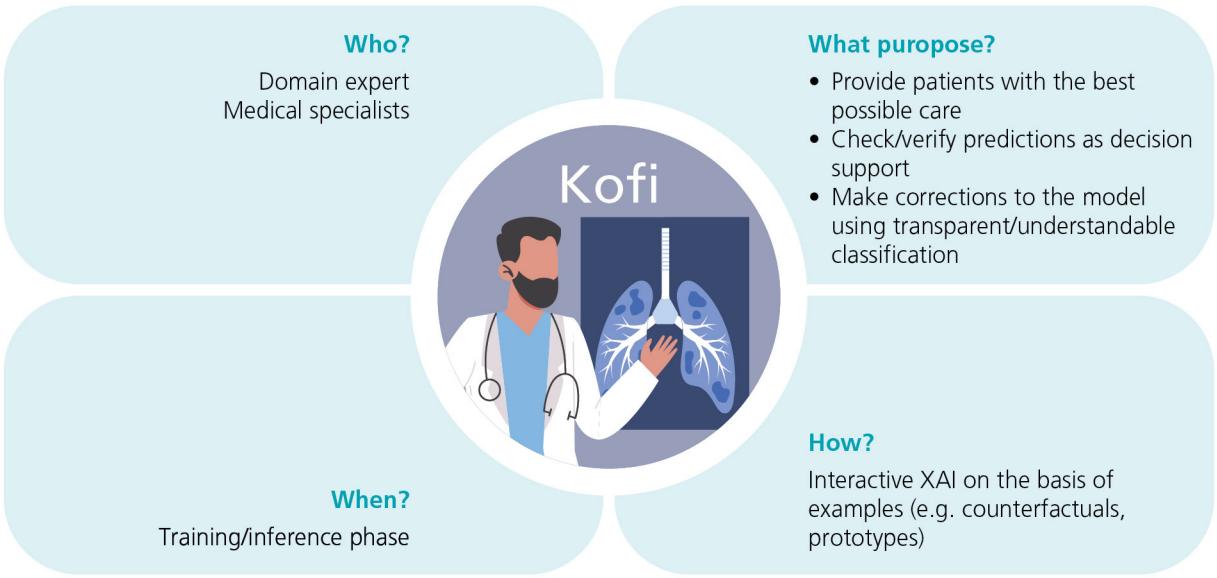
Concept-based

The XAI component of the AI system provides users with explanation(s) based on concepts, such as 'black wing colour' or based on the linking of concepts. Users recognise possible misclassifications or errors in the linking of concepts.

Source: concept based on Teso et al. (2023).

How an XAI-based explanation strategy can be designed from different perspectives along the four questions addressed above is illustrated using seven different personas: AI researcher, domain expert (chemistry or pharmaceutical research, medical specialist), AI model architect, software product manager, and end user. For each persona, overarching goals and specific explainability requirements are identified, and the corresponding XAI explanation approaches are described as in the following example:

Persona „Kofi – Domain Expert"



Outlook

XAI is a promising field that can significantly improve the quality and trustworthiness of AI systems. Due to its societal and economic relevance, the further development of XAI methods and their practical application are crucial to successfully address future challenges in AI development. Therefore, both general goals of XAI should be pursued equally: the use of XAI to achieve trustworthy AI and the use of XAI as an engineering tool to improve models.

To better realize the potential of XAI, adaptability and alignment with goals, target groups, and domains should be actively pursued and increasingly considered. Humans should always remain at the center to enable actionable insights. Evaluation frameworks should be developed that are generally applicable across studies, contexts, and settings. In research, established methods should be improved and XAI methods should be further developed for new types of AI, particularly with regard to holistic approaches that integrate the data, model, and output levels, as well as approaches focused on inspectability

and controllability of large models. In higher education, XAI should be more firmly embedded as an engineering tool within AI engineering in AI and data science programs. Companies and organizations should consider the implementation of XAI as an integral part of their AI strategy and increasingly rely on it to, for example, improve products, reduce internal communication barriers, or differentiate themselves from competitors.

Imprint
Editor: Plattform Lernende Systeme – Germany's Platform for Artificial Intelligence | Managing Office | c/o acatech | Karolinenplatz 4 | D-80333 Munich | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de | Follow us on LinkedIn: de.linkedin.com/company/plattform-lernende-systeme | Mastodon: social.bund.de/@LernendeSysteme | Status: June 2025 | Photo credit: janiecbros/iStock/Title

This executive summary is based on the white paper *Nachvollziehbare KI. Erklären, für wen, was und wofür*, Munich, 2025. The authors are members of the Working Group Technological Enablers and Data Science. The original version of the publication is available online at: *https://doi.org/10.48669/pls_2025-2*

Sponsored by the

Federal Ministry of Research, Technology and Space

acatech
NATIONAL ACADEMY OF SCIENCE AND ENGINEERING