

Explainable AI for EU AI Act compliance audits

Vincent Damen, Menno Wiersma, Gokce Aydin, Rens van Haasteren

Received 14 February 2025 | Accepted 14 July 2025 | Published 11 September 2025

Abstract

Internal auditors play a key role in ensuring artificial intelligence (AI) compliance with the EU AI Act. This article will examine how Explainable AI (XAI) can play a critical role in assessing AI systems for meeting the specific requirements of transparency, human oversight, and fairness. When effectively implemented, XAI enables traceability, accountability, intervention in AI decisions and can be used as a tool by internal auditors. Effective AI compliance auditing requires understanding of the methods for AI monitoring, associated documentation, and user feedback mechanisms to assess risks, regulatory requirements, and ethical standards.

Relevance to practice

While the internal audit function in the oversight of AI systems is not mandatory under the EU AI Act, their contribution to ensuring compliance with it is increasingly recognized as essential. Internal auditors can assess whether XAI layers added to AI systems sufficiently address transparency, human oversight, and fairness requirements. XAI supports traceability and accountability, enabling effective risk evaluation.

Keywords

Artificial intelligence, internal audit, EU AI Act, Explainable AI, transparency, human oversight, fairness

1. Introduction

High-risk applications of artificial intelligence (AI) underscore the critical need for reliable, accountable, and transparent AI systems. A clear example is found in credit risk assessment, where AI systems are used to determine whether individuals are eligible for loans. These decisions can significantly impact people's lives, making it essential that such systems are explainable and subject to human oversight. Under the European Union's Artificial Intelligence Act (EU AI Act), credit risk scoring is classified as high-risk and is therefore subject to strict transparency and accountability requirements. Actually, as this article will expand upon, all AI systems will need to (indirectly) adhere to some form of explainability requirements due to the EU AI Act.

The fundamental question of this article tries to resolve is:

“Can an explainability layer help AI deployers comply with the EU AI Act's transparency and oversight requirements, and how can internal auditors use it for compliance verification?”

The answer: it depends. Explainable AI (XAI) can support compliance by making model decisions more transparent and understandable. However, its effectiveness varies, as some methods oversimplify complex models or provide inconsistent interpretations. To be useful for internal auditors, explanations must be clear, reliable, and actionable, ensuring internal auditors can effectively assess compliance.

In a previous MAB article, we provided guidance how internal auditors can build a framework to audit AI

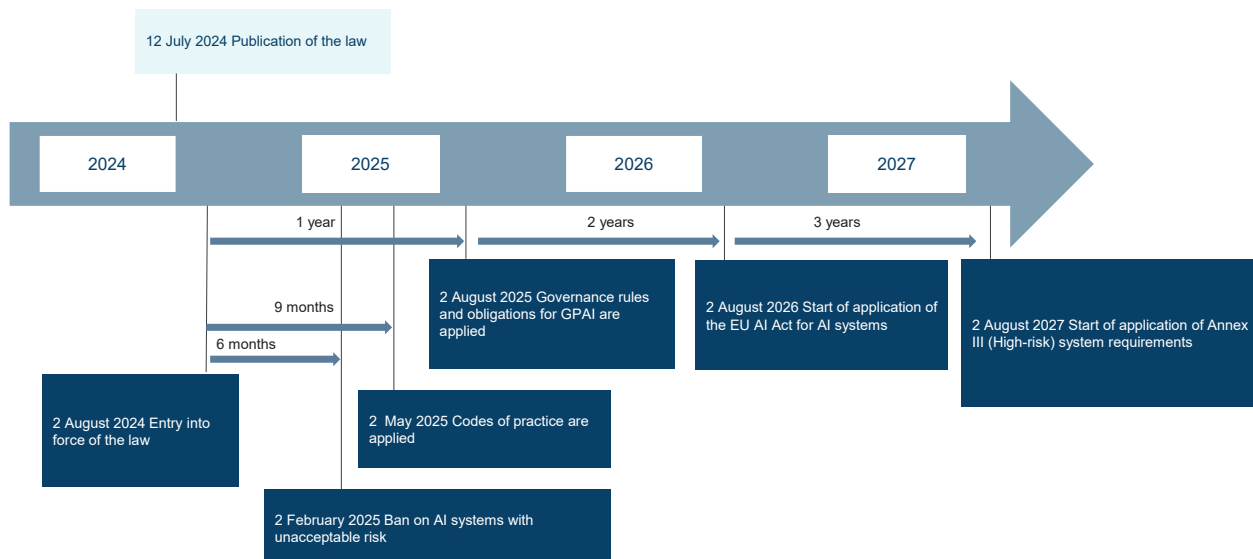
systems (Sandu et al. 2022), also taking into account the EU AI Act. In this article we zoom in on how XAI can play a critical role for the internal auditor in assessing the specific requirements of transparency, human oversight, and fairness. To understand the role of internal auditors within the EU AI Act, Chapter 2 will explore the structure of this new legal framework and detail articles related to the explainability and transparency of AI systems. Chapter 3 will discuss XAI and its limitations, demonstrated with a credit risk example. Chapter 4 will revisit the EU AI Act, now analyzed through the lens of an internal auditor, incorporating insights from the strengths and weaknesses of XAI. Finally, Chapter 5 will summarize the key concepts presented throughout the article.

2. EU AI Act (Transparency & Human Oversight Requirements)

2.1. Overview of the EU AI Act (parties and classification of systems)

The European Regulation EU 2024/1689 came into effect on June 13, 2024 (EP 2024). It is better known as the European Union Artificial Intelligence Act (EU AI Act) and it ranks as one of the first pieces of legislation attempting to regulate AI technologies. While the Act has already entered law, not all requirements are yet enforced and the roadmap for implementation extends until 2027, the timeline is shown in Figure 1.

Figure 1. EU AI Act Timeline.



Risk-based approach

Core to the EU AI Act is a risk-based classification system for AI-systems that have specific compliance requirements attached to them. The risk-based classification

system for AI technologies introduced in the Act, categorizes AI systems according to their potential impact on health, safety, fundamental rights and emphasizes transparency and human oversight (EP 2024, Article 6), illustrated in Figure 2. It is therefore essential to prohibit certain unacceptable AI practices, establish requirements for high-risk AI systems, and impose obligations on the relevant operators, while also setting transparency requirements for specific AI systems.


Roles in AI systems in the EU AI Act

Key operators in the AI value chain have been defined by the EU's regulatory framework, and are important to understand, as they determine compliance requirements based on role (EP 2024, Article 2). **Providers** are defined as entities that develop an AI system or place a general-purpose model on the market or put the AI system into service under its own name or trademark, whether for payment or free of charge. **Deployers** are the entities that use an AI system under their authority, except for natural persons using AI systems for personal, non-professional purposes. Finally, **an affected person** is an individual that uses or is affected by AI systems.

The following three sections of this chapter focuses on transparency requirements, human oversight, and fairness. Transparency requirements within the EU AI Act necessitate that AI systems provide clear and understandable explanations for their decisions, thus calling for the use of an explainability layer to make AI systems' decision-making processes more transparent. Human oversight ensures that AI systems are non-

itored based on model performance and explanations and can be corrected when necessary. Fairness principles seek to prevent biases in AI decision-making, and these biases can be identified and addressed using XAI techniques.

Figure 2. Risk-based classification system for AI systems.

| | |
|---|--|
|  | Prohibited AI Practices: (REF1, Article 5): These are AI systems or uses that pose unacceptable risks and are therefore outright prohibited. This might include manipulative AI practices that exploit vulnerabilities or deploy subliminal techniques. |
| | High-Risk AI Systems: (REF1, Article 6): These are AI systems, that can be also safety components of products or systems, or which are themselves products, identified as having significant potential to cause harm significant consequences for people's health, safety, or fundamental rights including financial exclusion and discrimination. High-risk AI systems should only be placed on the Union market, put into service or used if they comply with certain mandatory requirements. |
| | General-Purpose AI Models with Systemic Risk: (REF1, Article 51): These are AI models that have high-impact capabilities or an impact equivalent to those capabilities. They present systemic risks due to their reach or high-impact capabilities, which are evaluated using appropriate technical tools and methodologies. AI systems falling under this category must be transparent, meaning individuals must be informed when they are interacting with AI systems, unless this is self-evident. |
| | Limited-Risk AI Systems: While not explicitly defined in the EU AI Act, these are AI systems that may not fall under the high-risk or General-Purpose AI Models with Systemic Risk categories and are mandated to adhere to the general principles in the Act. These general principles equate to; Transparency, Human Oversight, Data Protection and Privacy, Accountability and Non-Discrimination. |
| | Low/Minimal Risk Systems: While not explicitly defined in the EU AI Act, this category includes AI systems that do not fall under any of the other categories. These systems, however, will also fall under the general principles and are not outside of the Act. |

2.2. Transparency requirements in the EU AI Act

The EU AI act is linked to the European Union's General Data Protection Regulation (GDPR), which addresses concerns about the opacity of decision-making processes by automated systems. The GDPR includes provisions for automated decision-making (ADM) based on personal data, establishing protections and safeguards for individuals when subjected to decisions based solely on automated processing. (EP 2016, Article 22). GDPR clarifies that data subjects are provided with meaningful information about the logic involved in ADM (EP 2016, Article 13, 14). Yet, the regulation does not offer clear guidelines on how such information should be conveyed in the context of ADM systems, leaving room for interpretation and inconsistency in implementation (Wörsdörfer 2024). In the context of the EU AI Act, transparency is underpinned by guided and structured rules, delineating to the obligations of AI system providers and deployers to disclose relevant information about the functioning, capabilities, and limitations of the AI systems they deploy or provide.

Risk-based requirements

In the EU AI Act, AI systems classified as High-risk require providers to supply comprehensive technical documentation. This includes a general description, intended use, and technical details such as system interaction with other hardware or software, and data used for training, including type and relevance (EP 2024, Article 13). The documentation should also be understandable to deployers who may not have specialist knowledge in AI, ensuring they are fully informed of the system's capabilities and limitations.

AI systems that interact with natural persons must disclose their AI nature unless it is obvious (EP 2024, Article 50). Additionally, non-high-risk systems, including

General-Purpose AI, must be clearly marked and labeled, ensuring users can distinguish AI from human interaction in consumer-facing applications.

Complaint mechanism

The EU AI Act strengthens transparency by introducing complaint mechanisms and a right to explanation, allowing individuals affected by AI-driven decisions to seek redress and clarification. This complements the obligation to inform users when they are interacting with an AI system. Under Article 85 (EP 2024), any natural or legal person may file a complaint with a market surveillance authority for suspected violations of the Regulation. Importantly, this mechanism applies to all AI systems, not just high-risk ones. The EU AI act also includes provisions for a right to explanation for decisions made using high-risk AI systems listed in Annex III (with certain exceptions). It states that affected persons have the right to obtain clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken (EP 2024, Article 86).

The complaint mechanism and the right to explanation create a need for explainability in AI decision-making, even though many models are too complex to provide clear justifications. As a result, it can be argued that the EU AI Act implicitly instructs the use of Explainable AI (XAI) techniques to ensure that AI decisions can be understood and communicated.

2.3. Human oversight requirements in the EU AI Act

The human oversight mechanism was first established under the GDPR, granting data subjects the right not to be subjected to solely automated decisions involving the processing of personal data that result in legal or similarly significant effects. In cases where such decisions are made, appropriate human supervision and intervention –

often referred to as the ‘human-in-the-loop’ effect – are required to safeguard fundamental rights and freedoms (Fügener et al. 2021).

Building on this foundation, the EU AI Act states that high-risk AI systems must be designed to enable effective human oversight. This includes mechanisms that allow humans to understand, monitor, and control the operations of AI systems. The objective is to ensure that AI systems are not autonomous but operate under human governance, enabling necessary interventions and informed decisions (EP 2024, Article 14).

Human oversight also extends to ethical and legal considerations. AI systems must be developed and operated in ways that respect fundamental rights and comply with applicable laws (EP 2024, Article 3). Human overseers play a critical role in ensuring that AI systems do not perpetuate biases or make decisions that could lead to discrimination or other ethical issues.

Development phase

During the design and development phases of AI systems, features facilitating human oversight must be incorporated into AI systems (EP 2024, Article 14). This involves creating interfaces or tools that allow humans to interpret the system’s outputs effectively. The goal is to maintain human control over the AI system, preventing scenarios where the system operates autonomously without human input or correction.

Monitoring

The Act further mandates that providers must implement measures for continuous monitoring of high-risk AI systems’ operation (EP 2024, Article 14). This monitoring should include mechanisms for detecting and responding to anomalies or unintended behaviors. Humans in oversight roles must have the ability and authority to intervene in real-time to correct or disable the AI system if it behaves unpredictably or deviates from its intended function.

Exemption

Moreover, the EU AI Act outlines low-risk scenarios, stating that an AI system that does not materially influence the outcome of decision-making should be understood as one that does not affect the substance or result of a decision, whether human or automated (EP 2024, Recital 53). Such systems can be considered exempt from stringent oversight mechanisms as they do not exert a substantial influence on decision-making outcomes.

It is argued that the “human-in-the-loop” model under GDPR and “human-oversight-by-design” under the EU AI Act focuses on ensuring human involvement, but do not define how to ensure that oversight is effective or competent (Laux 2023). Moreover, the obligations placed on AI developers to ensure proper oversight remains underdefined, leaving room for significant interpretation.

2.4. Fairness principle under the EU AI Act

AI systems are required to comply with the fundamental principles outlined in the EU AI Act. Among these principles, fairness is a cornerstone of trustworthy AI, ensuring that automated decision-making processes do not perpetuate bias or discrimination (EP 2024, Recital 27). However, despite its recognized importance, the EU AI Act does not provide explicit provisions on how fairness should be maintained. There is no clear obligation imposed on AI providers or deployers to assess, mitigate, or rectify model biases, leaving a regulatory gap.

While the EU AI Act does not have a single “Fairness” article, fairness principles are embedded in multiple provisions, particularly those related to bias mitigation (EP 2024, Article 10), human oversight (EP 2024, Article 14), transparency (EP 2024, Article 52), and fundamental rights assessments (EP 2024, Article 28).

The primary legal mechanism ensuring fairness is the requirement that high-risk AI systems must not result in discriminatory, biased, or unfair outcomes. Article 10 of the EU AI Act ensures fairness by requiring high-quality, bias-free, and representative training data for high-risk AI systems. The next chapter, we will explore the implications of XAI techniques on ensuring fairness in AI systems.

3. Explainable AI (XAI)

3.1. What is XAI?

AI is flexible in the way that it transforms raw information (input data) into the model’s prediction (output data) by finding the best statistical fit to ensure that the model captures the patterns in the data. Drawback is that it is not always directly clear what the exact relationship is, and why it is how it is. Such application is seen as ‘black box’ to the developer and the user, not knowing what happens inside. Users in this context are defined as IT users within the deployer of the AI system. High-risk AI systems must be sufficiently transparent to enable deployers to interpret their output and provide information that is relevant to explain or interpret their output, as suggested by the EU AI Act (EP 2024, Article 13). As an additional control to use a system appropriately and to ensure accuracy of output, human oversight needs to be established as it provides critical judgement and validation of predictions.

An explainability layer on top of an AI system, referred to as XAI, helps a user performing human oversight in the explanation and interpretation of the output. This is only interesting when the system uses black-box models. This can also be the case when using third-party tooling with proprietary models. Under the EU AI Act, it is not sufficient to solely explain the overall functioning of the AI system (global), the output for a specific input needs to be also explained (local), as the right to complain requires a local explanation. The XAI explainability

layer and techniques are not only important for ensuring fairness, but might also be interesting for organizations seeking to comply with both the transparency and human oversight requirements of the EU AI Act.

3.2. XAI characteristics

In addition to a technique giving global or local explanations, and the type of input data that the XAI technique can work with, there are other aspects that characterize specific XAI techniques, which should be considered when designing the aforementioned explainability layer:

- Some XAI techniques rely on an *assumption of independence between features* (explanatory variable), which leads to the omission of interactions between those features. This assumption can oversimplify real-world scenarios, potentially compromising the accuracy and relevance of the insights provided.
- The *implementation difficulty* of such techniques encompasses the complexity and time required to create an effective explanation layer. Techniques that require developers to make intricate implementation decisions or finely tune parameters are considered hard to implement. This is a key metric for validators or internal auditors, whose proficiency vary significantly, as programming is not their core skill set.
- Another critical aspect of an XAI technique is its *clarity*, i.e. user-friendliness (Gerlings et al. 2020; Y and Challa 2023). Techniques that produce multiple subfigures, use non-standard axis or present layered visuals are regarded as having low clarity.
- Closely related is the concept of *information density*, which covers how much information is conveyed. Techniques that only reveal a single dimension of insight or provide superficial details are considered to have low information density.
- Finally, *computational complexity* is the amount of resources required to run it (Y and Challa 2023). It illustrates the time it takes to run the XAI technique. Higher computational complexity results in longer processing times, which could impact the feasibility in environments with limited computational resources.

3.3. XAI design considerations

Panigutti et al. (2023) describe that the model can also be transparent by design, leading to interpretable AI (IAI). The decision-making process can then be directly assessed by internal auditors and validators because of the simple understandability of the inner workings of the

model. Examples of IAI are relatively simple linear, tree-based, rule-based models, sparse models or models that process information in a way that is interpretable (Panigutti et al. 2023).

The explainability of an outcome cannot be looked at on a standalone basis. It needs to be assessed together with the performance of the model and stability of that performance. Low accuracy and/or stability need to be reflected in the explanation, to get a full understanding of the relationship and how strong it is.

In addition, the way of implementing explainability needs to be suitable to the level of expertise of the users performing oversight, and the users need to be sufficiently trained. In many cases, this may require additional representation tooling on top of the core XAI techniques. Based on a literature study, Haque et al. (2023) describe users' explanation needs and the effect of explanation on users' perceptions of an AI system. Dependent on the type of users, different representation formats may be used. The representation needs to be complete and sufficiently accurate for well-informed decision-making. The perception is influenced by the level of communicated versus observed accuracy, transparency of the working of the system, understandability due to sufficient involvement, added value of the explainability, and perceived fairness of outcomes with especially local explainability.

3.4. Example model

To illustrate the benefits and limitations of XAI, a specific banking application is developed in this section to estimate if a specific loan is likely to default or not, based on a set of features. The section afterwards demonstrates several XAI techniques given this credit risk model. Credit risk models are high-risk models under the EU AI Act, and as such the risk classification must be explainable due to the transparency requirements of the EU AI Act.

Ferreira (2018) selected some important features from a dataset published by Hofmann (1994). The variable to predict is the risk, where 'good' and 'bad' signify no default and default, respectively. The table of (Ferreira 2018) is shown in Table 1.

After encoding the ordinal (ordered variables: 'Saving account', 'Checking account') and nominal (unordered variables: 'Sex', 'Housing', 'Purpose') variables, the data is randomly split into 80% training and 20% test data. A Random Forest Classifier (introduced by Breiman (2001)) has been used to predict the risk variable, this could have also been any other classification model. This model results in an accuracy of 0.74. The results from the predictions are shown in Table 2, where on the columns

Table 1. First two rows of example dataset.

| Age | Sex | Job | Housing | Saving account | Checking account | Credit amount | Duration | Purpose | Risk |
|-----|--------|---------|---------|----------------|------------------|---------------|----------|----------|------|
| 67 | Male | Skilled | Own | None | Little | 1169 | 6 | Radio/tv | Good |
| 22 | Female | Skilled | Own | Little | Moderate | 5951 | 48 | Radio/tv | Bad |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

Table 2. Prediction results.

| Ground-truth/prediction | Predict as ‘good’ by model | Predict as ‘bad’ by model | Total actual |
|-------------------------|----------------------------|---------------------------|--------------|
| Actual good | 127 (63,5%) | 38 (19%) | 165 (82,5%) |
| Actual bad | 14 (7%) | 21 (10,5%) | 35 (17,5%) |
| Total predicted | 141 (70,5%) | 59 (29,5%) | 200 (100%) |

the prediction of the model is placed and on the rows the ground-truth. For example, out of the 200 credit loans of which the risk level was predicted, the model classified 38 credit loans as having a ‘bad’ risk, while they were actually loans with a good risk profile.

3.5. XAI techniques

Two of the most widely used XAI techniques are LIME and SHAP. LIME provides a local explanation and SHAP can give a local and a global explanation.

Local Interpretable Model-agnostic Explanations (LIME)

LIME uses a simple model to approximate a complex model (Ribeiro et al. 2016). It tries to explain the output of an individual observation by creating a sample with slightly randomly changed inputs and looking at the outputs. A simple interpretable linear model is used to fit this sample and explain the output. A LIME analysis for

a specific instance (single loan application) of the credit risk example model is demonstrated in Figure 3. LIME states that this instance has an 82% probability of being a bad loan, seen on the left. In the middle graph it is shown that a low checking account and a low age are the main drivers for a bad loan, whereas the features with the blue coloring are drivers for a good loan. On the right, the value for each feature is shown.

SHapley Additive exPlanations (SHAP)

SHAP uses Shapley values that define the importance of an individual explanatory variable (feature), as the relative change in the output, with the specific feature included versus when it is excluded (Lundberg and Lee 2017). A SHAP plot is given for a specific instance, which is demonstrated in Figure 4 (left). A global SHAP explanation of the model is demonstrated in Figure 4 (right), where each dot is an instance of the data and the feature value coloring is the raw relative value of the feature.

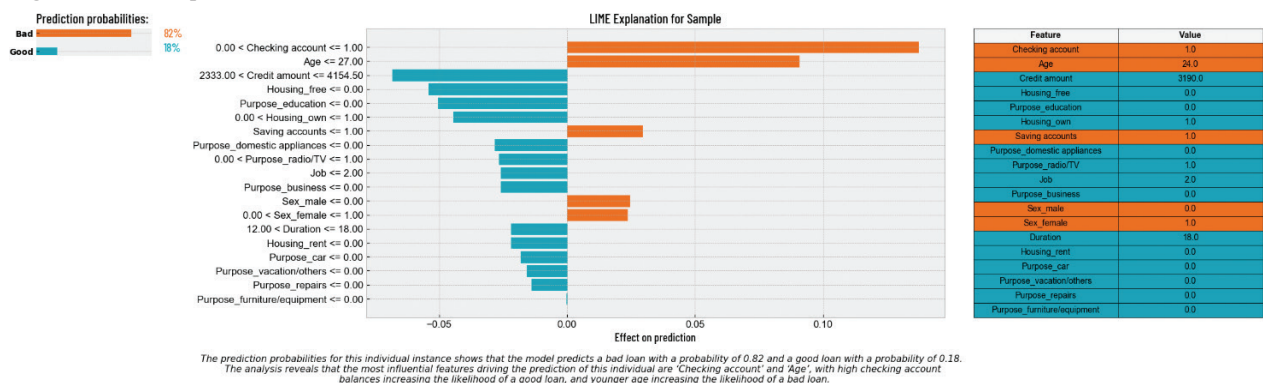
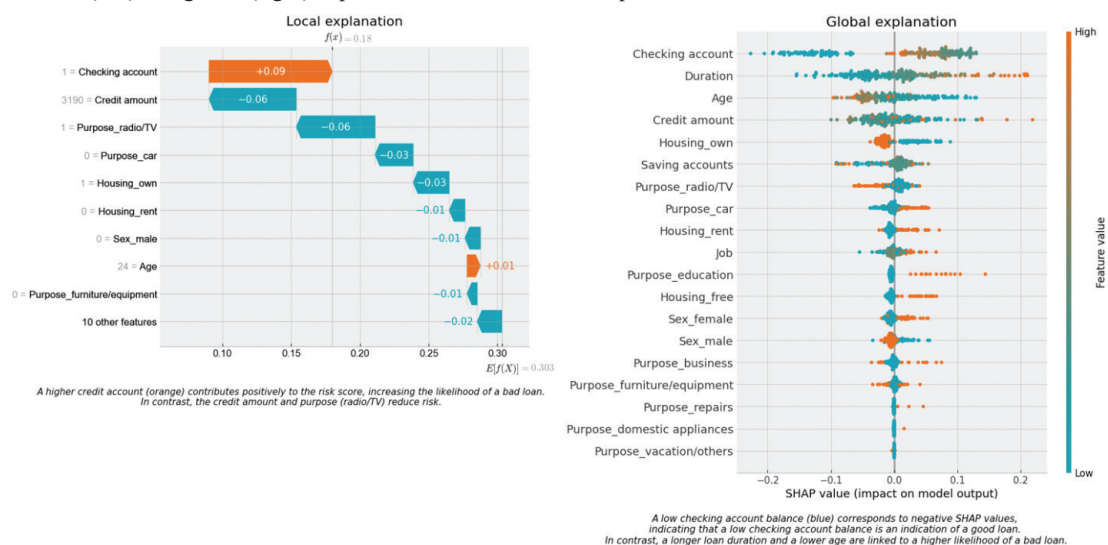
Figure 3. Local explanation from LIME.**Figure 4.** Local (left) and global (right) explanation from SHAP technique.

Table 3. Overview of XAI techniques.

| XAI technique | Objective | Type of input data | Global vs local | Assumes independence of features | Implementation difficulty | Clarity | Information density | Computational complexity | Remarks |
|--|--|--------------------|-----------------|----------------------------------|---------------------------|---------|---------------------|--------------------------|---|
| LIME | Explaining individual predictions | Text Tabular Image | Local | No | Easy | Medium | High | Medium | Struggles with high-dimensional data |
| SHAP | Understanding global feature importance | Text Tabular Image | Both | No | Medium | Low | High | High | Mathematically grounded, based on cooperative game theory |
| Global surrogate model | Summarizing complex models | Tabular | Global | No | Medium | Medium | Medium | Medium | May oversimplify complex models |
| Anchors method | Explaining rule-based models | Tabular | Local | No | Hard | High | Medium | High | Struggles with high-dimensional data. |
| Counterfactual explanation | Identifying actionable changes to outcomes | Text Tabular Image | Local | No | Hard | Low | Medium | High | Difficult to find useful explanations for high-dimensional data |
| Permutation Feature Importance (PFI) | Assessing feature importance | Tabular Image | Global | No | Easy | High | Low | Low | Assumes independence between features |
| Partial Dependence Plot (PDP) | Visualizing relationships between features and predictions | Tabular | Global | Yes | Easy | High | Low | Low | Suitable for uncovering average effects |
| Individual Conditional Expectation (ICE) | Exploring feature impact on specific instances | Tabular | Both | Yes | Easy | Medium | Medium | Low | Lacks scalability |
| Accumulated Local Effects (ALE) plot | Improving global feature analysis | Tabular | Global | Yes | Easy | Medium | Medium | Medium | Suitable for identifying local effects in correlated datasets |
| Friedman's H-statistic | Detect interactions between features | Tabular | Global | No | Easy | Easy | Medium | High | Has underlying theory from partial dependency decomposition |
| MDD-critic | Identify representative and not representative datapoints | Tabular | Global | No | Medium | High | Medium | High | Difficult to select proper number of prototypes and criticisms |

3.6. Overview

Table 3 presents the mentioned LIME and SHAP but also several other XAI techniques and their characteristics. These techniques are selected in their ability for assessing transparency, human oversight and fairness (Molnar 2019; Zhang et al. 2022). The characteristics (implementation difficulty, clarity, information density and computational complexity) are ranked relative to each other.

All these techniques are model agnostic, although there are variants that are model-specific. This means that XAI is quite flexible. Different techniques, or a combination of techniques may be used, dependent on the type of input data, sophistication of the XAI developer, sophistication of the user of XAI, and accuracy and consistency required.

Ease of implementation

The techniques are well available in standard or specific libraries of statistical programming languages such as Python and R, so that with limited effort an explainable layer can be added to an AI system. It may be beneficial in a validation where explainability was not embedded during development (e.g. for low risk systems), to implement XAI to get better understanding about the working

of a model and where risks manifest (Zhang et al. 2022). The most applicable packages for each XAI technique are shown in Table 4.

Table 4. Python and R packages for several XAI techniques.

| XAI technique | Python package | R package |
|--|-------------------|-----------------|
| LIME | Lime | Lime |
| SHAP | Shap | Shapr |
| Global Surrogate Model | Scikit tree | Iml |
| Anchors method | Alibi | Party |
| Counterfactual explanation | DiCE | Counterfactuals |
| Permutation Feature Importance (PFI) | Scikit inspection | Vip, iml, DALEX |
| Partial Dependence Plot (PDP) | Scikit inspection | Pdp, iml, DALEX |
| Individual Conditional Expectation (ICE) | Scikit inspection | Ice, iml, pdp |
| Accumulated Local Effects (ALE) plot | PyALE | ALEPlot |
| Friedman's H-statistic | Artemis | Iml |
| MDD-critic | mmd-critic | eummd |

3.7. Limitations

While XAI offers significant benefits when tailored to stakeholders' needs, it also comes with notable limitations. Therefore, just like with AI itself, XAI cannot be imple-

mented as a tool that will automatically resolve all transparency, human oversight, and fairness issues. Expert involvement is essential in choosing how to apply XAI, what method(s) to use, how to interpret results, how to communicate these, and to opine on the AI system in areas where XAI was not applied. Most important limitations are:

- **Judicial.** XAI does inherently not guarantee AI decisions are legitimate, reliable, or unbiased.
- **Inconsistency and irreproducibility.** Different XAI methods can yield varying results for the same model. Some, like LIME and Anchors, introduce randomness, making explanations unstable.
- **Automation bias.** XAI can create a false sense of reliability, known as automation bias, leading to reduced human oversight and errors in decision-making going unnoticed (Schemmer et al. 2022). Underlined by the credit risk models, which might make significant errors if the operators rely too much on the decision making of the risk model without understanding the underlying model, showing the crucial importance of human judgement.
- **Fairness concerns.** As mentioned in the previous section, the EU AI Act emphasizes fairness as a key principle (EP 2024, Recital 27). While XAI can help reveal biases (McDermid et al. 2021; Hofeditz et al. 2022; Chuan et al. 2024), it does not inherently ensure fair decision-making. Analysis highlights that biases can persist through proxy variables; simply removing the ‘Sex’ feature from the credit risk example will not directly make the model fair. Ensuring fairness requires careful design, stakeholder involvement, and transparency behind XAI techniques (Longo et al. 2024). In this context, Dwork et al. (2012) introduced the concept of *Fairness Through Awareness*, which emphasizes that fairness should be understood and addressed relative to the specific context of the decision-making process.
- **Selection of XAI techniques.** The effectiveness of XAI depends on the chosen methods, which vary by stakeholder needs. A one-size-fits-all approach can cause misinterpretation. XAI should adapt to user needs and include educational tools. Developers may also manipulate representations to hide biases (Deck et al. 2023).

4. What does it mean for the internal auditor?

4.1. The role of the internal auditor

Internal auditors will play a critical role in evaluating AI systems in the context of meeting the transparency, human-oversight and fairness requirements of the EU AI Act. See also our previous article (Sandu et al. 2022) including a proposed framework for auditing algorithms in general, and the life cycle approach for continuous

involvement of the internal audit department. As defined in Chapter 1, while the involvement of Internal Audit (IA) professionals in overseeing AI systems is not mandated by the EU AI Act, the IIA’s publication on the AI Act (IIA 2023) underscores their vital role in assessing AI risks, promoting transparency, and ensuring that governance frameworks align with regulatory expectations. The publication particularly highlights two critical contributions of the IA function:

- Advisory Capacity* – Internal auditors support management by providing guidance on how AI should be effectively managed, developed, and governed.
- Assurance Function* – Internal auditors independently assess whether AI-related controls and processes are properly designed, implemented, and functioning as intended.

Advisory capacity

While the EU AI Act does not explicitly assign responsibilities to internal auditors, it imposes clear compliance and documentation obligations on deployers of high-risk AI systems. Deployers are required to conduct a fundamental rights impact assessment (EP 2024, Article 27), maintain up-to-date documentation (EP 2024, Annex C), and cooperate with competent authorities in enforcement actions (EP 2024, Article 27(12)). These responsibilities naturally fall within the scope of internal audit functions, which are typically tasked with providing independent assurance over regulatory compliance, risk management, and control effectiveness. As such, internal auditors are well-positioned to provide assurance that the deployer’s obligations under the EU AI Act are being fulfilled.

Assurance function

Considering internal auditors will be tasked to provide assurance that the outputs of AI systems can be understood and explained, not just for their functionality, but also to verify adherence to fundamental rights, safety, and ethical principles as mandated by the EU AI Act (ECIIA 2024), we must define what assurance means for the internal auditor in this context. According to the IIA’s Global Internal Audit Standards (IIA 2024), assurance is a statement intended to increase the level of stakeholders’ confidence about an organization’s governance, risk management, and control processes over an issue, condition, subject matter, or activity under review (e.g. AI system compliance with the EU AI Act requirements) when compared to established criteria. Internal auditors may provide limited or reasonable assurance, depending on the nature, timing, and extent of procedures performed.

The EU AI Act also does not mandate a specific assurance reporting format for providing assurance on AI systems. In order to have a structured approach for AI assurance and reporting purposes, internal auditors could for example align with the International Standard on Assurance Engagements 3000 (ISAE3000) (IAASB

2024). The ISAE 3000 offers a common methodology for both internal and external auditors such as chartered accountants or IT-auditors to structure their assurance work. Internal auditors may adopt ISAE 3000 principles to guide their evaluations, while external auditors can collaborate with the internal audit function by validating or supplementing the internal audit's findings.

In the context of AI, internal auditors would apply assurance by evaluating whether the AI system complies with the EU AI Act's requirements. Assessing the design and effectiveness of internal controls related to AI governance, transparency and oversight, and reviewing documentation such as risk assessments, logs, and human oversight protocols. If an AI system lacks these capabilities and has not been challenged by a second line, internal auditors may need to assess it by implementing an explainability layer themselves. This may be the case when second-line challenge is not available (especially in a non-financial institution environment). In such a case the internal auditor takes up a combined second- and third-line role. If assessing an already implemented and used AI system, it may not be sufficient for the internal auditor to notice a design deficiency in case an explainability layer is missing. If already implemented, the internal auditor may also want to test operational effectiveness by independently adding an explainability layer to the system. It is likely that external expertise would be required to support the internal audit function. Chapter 3 helps to understand the capabilities to look for when hiring external expertise. In case there is a second line function available, it is more likely that the internal auditor feeds back to the first and second line to resolve.

In case explainability is part of the system, internal auditors must evaluate whether this explainability is sufficient for the system's intended use and matches users' level of understanding. Moreover, internal auditors should evaluate whether the level of explainability is adequate for the system's intended use and aligns with users' ability to interpret it. As highlighted by the European Confederation of Institutes of Internal Auditing (ECIIA), it is considered good practice for organizations to establish policies, procedures, or guidelines that define explainable AI (XAI) requirements and their practical implementation to support compliance efforts (ECIIA 2024).

4.2. The role of XAI in assessing transparency and human oversight

As AI systems become more integral to organizational decision-making, Explainable AI (XAI) serves as a key mechanism for internal auditors to evaluate whether these systems comply with the EU AI Act. Specifically, XAI supports critical assessments of transparency, human oversight and fairness, which are central obligations under Articles 10, 13, and 14 of the EU AI Act (EP 2024) that were addressed previously in Chapter 2.

XAI enhances transparency by making AI decision-making processes understandable to human stakeholders. For

internal auditors, this involves ensuring that the decision logic behind AI outputs can be clearly articulated, verifying whether input features, processing steps, and model decisions are documented and interpretable, assessing whether users can understand how outcomes are generated.

XAI also plays a role in evaluating fairness, particularly in ensuring that AI systems do not produce discriminatory outcomes. Internal auditors should apply XAI to detect bias patterns in both training data and model logic.

By enabling transparent inspection of model behavior, XAI also supports compliance with Article 10 on data quality and governance and helps uphold broader EU human rights and non-discrimination principles. According to Article 14, AI systems must allow for meaningful human oversight, enabling human intervention where necessary. XAI contributes to this by providing actionable explanations that support human operators in overriding or correcting AI outputs.

Internal auditors can use these insights to evaluate whether oversight mechanisms are not just formally present but also functionally effective. It is thus considered good practice for an organization to implement policies, procedures, or guidelines outlining XAI requirements and their application to enable this.

Transparency and explainability

From an EU AI Act compliance perspective, particularly concerning high-risk systems, as well as for risk management practices for systems with a different classification, outputs need to be understandable and explainable. This means that during the development phase, consideration should be given to either designing interpretability into the system or layering explainability on top of it. If explainability is absent, this indicates a design deficiency, potentially leading to biases and unwanted behavior. Depending on the risk and impact of the system, this may be a blocking issue, in which case there is no added value for the internal auditor for testing operational effectiveness. Alternatively, internal auditors may test operational effectiveness by adding an explainability layer during their review, e.g. also in case the system is already in use (see also Chapter 4). In either case, internal auditors must assess whether the appropriate XAI techniques are employed to satisfy explainability requirements. These requirements must be clearly defined, addressing characteristics such as transparency and user comprehension.

However, it is important to note that XAI does not directly equate to meeting the Transparency and Human Oversight requirements outlined in the EU AI Act. Transparency and Human Oversight entail broader considerations, such as ensuring meaningful human intervention and accountability at critical points in the AI lifecycle, as described in the EU AI Act. While XAI may enhance explainability, internal auditors should carefully evaluate whether the organization's approach to XAI truly addresses the EU AI Act's regulatory standards or if additional measures are needed to meet these obligations.

Human oversight

The potential non-compliance and liability risks associated with incorrect decisions made by AI systems, whether direct or indirect, underscores the need for substantiating why certain decisions were made by the system. This is where human oversight becomes essential. High-risk systems must incorporate human interaction within their processes. For systems classified differently, human oversight is a requirement when incorrect outcomes occur, and affected individuals require an explanation. In both scenarios, users need to understand the system's outputs and have the ability to provide localized explanations. Audit testing should ensure organizations have processes in place to uphold fundamental rights under the EU AI Act. Any model requires explainability, as the ability for customers to file a complaint with a market surveillance authority if they suspect a violation applies to all AI systems covered by the regulation, not just high-risk systems. The EU AI Act also gives individuals the right to an explanation for decisions made by high-risk AI systems listed in Annex III, with some exceptions. Affected individuals must receive clear explanations about the AI system's role in decision-making and the key factors influencing the outcome. To support this, organizations can use XAI to showcase transparency by providing insights into how AI systems operate. Organizations must also show that users, including those handling complaints, are properly trained to understand the system and its explainability features, ensuring compliance with the Regulation. For the internal auditor it is important to understand that different AI systems may deliver varying levels of accuracy or stability over time. To ensure a thorough understanding of the model, performance-related information should be a part of the explainability process. Well-designed XAI systems incorporate this into the explanations provided to users.

The internal auditor's responsibilities extend beyond the development phase. Ongoing monitoring and review of the AI system, guided by established policies, must include an assessment of the XAI layer's effectiveness. A user feedback loop should also be implemented, enabling users to consistently provide input on the system's performance, particularly regarding any malfunctions or areas for improvement. This feedback is essential for future system improvements, ensuring both functionality and explainability remain robust over time. Below, we break down key areas in greater detail to help internal auditors deliver impactful results.

4.3. Auditing AI systems leveraging XAI

Internal auditors evaluating an organization's ability to leverage XAI to meet the transparency, human oversight and fairness requirements outlined in the EU AI Act will require a structured approach to ensure compliance.

The first step would be to understand the organization's AI governance framework (see also our previous article, Sandu et al. (2022)). This involves examining

policies, procedures, and controls established to ensure responsible AI practices. Internal auditors should evaluate whether these frameworks address transparency and oversight, including the existence of XAI principles. Attention should be paid to documented processes for risk assessment, decision-making accountability, and alignment with the EU AI Act.

The internal auditor should also assess the technical capabilities of the AI system. This includes determining if the system provides understandable and accurate explanations for its decisions or outputs (the transparency). It needs to be assessed to what extent the explainability or interpretability of the model meets the standards required for transparency under the EU AI Act. Internal auditors should also assess the technical documentation provided with the system.

As we have seen, an important aspect of compliance is ensuring that human oversight mechanisms are in place. Internal auditors should assess whether human reviewers have the necessary tools, authority, and expertise to oversee AI decisions. This includes checking for procedures, workflows and tollgates that allow humans to intervene or override decisions made by the AI system in case of errors or ethical concerns.

In addition to transparency and human oversight, internal auditors should assess the fairness by reviewing records regarding the functioning of AI system, particularly those employing XAI. This includes logs of system decisions, interventions, and updates to the model or data. These records, in essence, provide evidence of compliant operations over a period of time, very useful when performing any sort of Test of Effectiveness (ToE) on the AI system.

AI standards and frameworks

Internal auditors should also compare the organization's practices with the guidelines, standards, best practices provided by regulatory bodies on transparency and human oversight in relation to XAI. Industry standards from ISO/IEC, NIST and of course the IIA can serve as a valuable reference for compliance evaluation.

One of the key tools for internal auditors to leverage on, is the IIA updated AI Auditing framework (The IIA 2023). The framework emphasizes the importance of audit trails, logs, and documentation as part of the internal control environment. These records, such as logs of system decisions, human interventions, and model/data updates, are explicitly recognized as critical for demonstrating accountability over time, supporting ToE procedures and enabling traceability of decisions, especially in high-risk or regulated environments.

Through the IIA AI Auditing Framework, internal auditors are encouraged to benchmark organizational AI practices against regulatory guidance (e.g., EU AI Act, U.S. Executive Orders) and industry standards such as ISO/IEC 22989 (AI Concepts and Terminology), ISO/IEC 23894 (AI Risk Management) and the NIST AI Risk

Management Framework (ISO 2022; ISO 2023; NIST 2023). Additionally, it recommends that auditors assess whether the organization has adopted transparency-enhancing practices such as clear documentation of model logic and limitations, human-in-the-loop mechanisms and explainability protocols for stakeholders.

While the IIA AI auditing framework does not prescribe a single method for XAI, it acknowledges the growing importance of explainability in AI governance. It suggests that internal auditors should evaluate whether the AI system provides meaningful explanations to users and stakeholders. They should also assess whether an XAI layer is documented, used appropriately and confirm that human oversight mechanisms are in place and effective.

5. Conclusion

In conclusion, this article makes it clear that XAI can play a crucial role in enabling internal auditors to assess compliance with the transparency, human oversight and fairness requirements outlined in the EU AI Act. For any kind of application, and any level of risk, in the design there needs to be a mechanism in place, by which the outcome of individual cases can be explained. The internal auditor needs to test if the design of the model is effective from that perspective and compliant with the EU AI Act. In case it is designed effectively, but also when it is designed ineffectively, XAI can equip internal auditors to test operating effectiveness of the core AI system (see Chapter 4). These design and operating effectiveness tests are fundamental to assessing adherence to the regulatory requirements of the EU AI Act.

One of the primary ways XAI supports internal auditors, is through its ability to produce detailed, human-readable explanations of AI-driven decisions. This feature ensures that internal auditors can trace the logic behind specific outcomes, identify potential biases or errors, and verify whether decisions align with the organization's ethical and operational objectives. Such transparency is critical for demonstrating compliance with the EU AI Act, which emphasizes accountability and the need for documented processes in the deployment of AI systems. Additionally, XAI's capacity to generate detailed logs, track system updates, and explain decision pathways makes the traceability and auditability of AI systems possible. These capabilities allow internal auditors to maintain a record of system operations, making it easier to evaluate changes over time and ensure ongoing alignment with regulatory frameworks.

Importantly, this article also lists important limitations to the use of XAI. In addition to explainability, the integration of human oversight mechanisms, as outlined in the EU AI Act, ensures organizations remain accountable. The incorporation of these mechanisms into XAI-supported processes enables protocols for intervention in cases of anomalies, errors, or decisions with potentially adverse consequences. Internal auditors can use XAI to identify these issues proactively, ensuring timely corrective actions are taken.

From a practical perspective, aligning XAI practices with established industry standards and frameworks, such as those provided by the IIA, ISO/IEC and NIST, internal auditors can ensure their processes are structured and are consistently supporting compliance assessments. This alignment not only supports internal auditors in validating AI system operations, but also enhances the credibility of their findings as they are based on industry best practices.

-
- **V.A. Damen RE, CISA – Vincent**, Associate Director Internal Audit & Financial Audit, Protiviti The Netherlands.
 - **Drs. M.R. Wiersma CFA, FRM, ERP – Menno**, Senior Manager Model Risk Management, Protiviti The Netherlands.
 - **G. Aydin LL.M., CIPM, CIPP/E, PRMIA – Gokce**, Operational Risk Certified, Senior Consultant Risk & Compliance, Protiviti The Netherlands.
 - **R. van Haasteren BSc – Rens**, Artificial Intelligence Intern, Protiviti The Netherlands.
-

References

- Breiman L (2001) Random Forests. *Machine Learning* 45: 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chuan CH, Sun R, Tian S, Tsai WHS (2024) EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives. *Telematics and Informatics* 91: 102135. <https://doi.org/10.1016/j.tele.2024.102135>
- Deck L, Schoeffer J, De-Arteaga M, Kühl N (2023) A critical survey on fairness benefits of XAI. XAI in Action: Past, Present, and Future Applications (preprint). <https://doi.org/10.1145/3630106.3658990>
- Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. <https://doi.org/10.1145/2090236.2090255>

- ECIIA (2024) The AI Act: Road to compliance. A Practical Guide for Internal Auditors. <https://www.eciia.eu/wp-content/uploads/2025/01/The-AI-Act-Road-to-Compliance-Final.pdf>
- EP (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>
- EP (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689
- Ferreira L (2018) German Credit Risk – With Target. Kaggle. <https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk>
- Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly* 45(3): 1527–1556. <https://ssrn.com/abstract=3879937>
- Gerlings J, Shollo A, Constantiou I (2020) Reviewing the need for explainable artificial intelligence (xAI). <https://doi.org/10.24251/HICSS.2021.156>
- Haque AKMB, Islam AKMN, Mikalef P (2023) Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change* 186: 122120. <https://doi.org/10.1016/j.techfore.2022.122120>
- Hofeditz L, Clausen S, Reiß A, Mirbabaie M, Stieglitz S (2022) Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets* 32(4): 2207–2233. <https://doi.org/10.1007/s12525-022-00600-9>
- Hofmann H (1994) Statlog (German Credit Data) [Dataset]. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>
- IAASB (2013) International Standard on Assurance Engagements (ISAE) 3000 Revised: Assurance engagements other than audits or reviews of historical financial information. [ISBN 978-1-60815-167-7] <https://www.iaasb.org/publications/international-standard-assurance-engagements-isae-3000-revised-assurance-engagements-other-audits-or>
- IAASB (2024) Handbook of international quality management, auditing, review, other assurance, and related services pronouncements. <https://www.iaasb.org/publications/2023-2024-handbook-international-quality-management-auditing-review-other-assurance-and-related>
- ISO (2022) ISO/IEC 22989:2022: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology.
- ISO (2023) ISO/IEC 23894:2023: Information technology — Artificial intelligence — Guidance on risk management
- Laux J (2023) Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & Society*, 1–14. <https://doi.org/10.1007/s00146-023-01777-z>
- Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser JD, Guidotti R, Hayashi Y, Herrera F, Holzinger A, Jiang R, Khosravi H, Lecue F, Malgieri G, Páez A, Samek W, Schneider J, Speith T, Stumpf S (2024) Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106: 102301. <https://doi.org/10.1016/j.inffus.2024.102301>
- Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *CoRR abs/1705.07874*. <https://doi.org/10.48550/arXiv.1705.07874>
- McDermid JA, Jia Y, Porter Z, Habli I (2021) Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A* 379(2207): 20200363. <https://doi.org/10.1098/rsta.2020.0363>
- Molnar C (2019) Interpretable machine learning. <https://christophm.github.io/interpretable-ml-book/>
- NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). <https://doi.org/10.6028/NIST.AI.100-1>
- Panigutti C, Hamon R, Hupont I, Llorca DF, Yela DF, Junklewitz H, Scalzo S, Mazzini G, Sanchez I, Garrido JS, Gomez E (2023) The role of explainable AI in the context of the AI Act. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 1139–1150. <https://doi.org/10.1145/3593013.3594069>
- Ribeiro MT, Singh S, Guestrin C (2016) “Why should I trust you?”: Explaining the predictions of any classifier. *CoRR abs/1602.04938*. <https://doi.org/10.18653/v1/N16-3020>
- Sandu I, Wiersma M, Manichand D (2022) Time to audit your AI algorithms. *Maandblad voor Accountancy en Bedrijfseconomie* 96(7/8): 253–265. <https://doi.org/10.5117/mab.96.90108>
- Schemmer M, Kühl N, Benz C, Satzger G (2022) On the influence of explainable AI on automation bias. <https://doi.org/10.48550/arXiv.2204.08859>
- The IIA (2023) The IIA’s Artificial Intelligence Auditing Framework. <https://www.theiia.org/en/content/tools/professional/2023/the-iias-updated-ai-auditing-framework/>
- The IIA (2024) Global Internal Audit Standards. <https://www.theiia.org/en/standards/2024-standards/global-internal-audit-standards/>
- Wörsdörfer M (2024) Mitigating the adverse effects of AI with the European Union’s artificial intelligence act: Hype or hope? *Global Business and Organizational Excellence* 43(3): 106–126. <https://doi.org/10.2139/ssrn.4630087>
- Y S, Challa M (2023) A comparative analysis of explainable AI techniques for enhanced model interpretability. In: *3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, Salem, India, 229–234. <https://doi.org/10.1109/ICPCSN58827.2023.00043>
- Zhang C, Cho S, Vasarhelyi M (2022) Explainable artificial intelligence (XAI) in auditing. *International Journal of Accounting Information Systems* 46: 100572. <https://doi.org/10.1016/j.accinf.2022.100572>