Examples of Scenarios for the Application of Explainable Artificial Intelligence in the Domain of Information Systems

Filip Krstić, dr Dušan Vujošević

***Abstract* - In an era of increasingly intensive implementation of large language models within information systems, the issue of explainable artificial intelligence (XAI) has become crucial for maintaining trust, transparency, and ethical accountability. This paper examines methodological approaches to XAI in the context of both traditional AI and LLM-based models, focusing on their applicability, limitations, and role in meeting regulatory and ethical requirements. Through theoretical analysis, comparative evaluation of post-hoc and intrinsic methods, and practical insights involving models such as DistilBERT and GPT-5, the study illustrates how various XAI techniques can contribute to a deeper understanding of the behavior of complex AI systems. Particular emphasis is placed on five scenario-based examples of the development and testing of solutions in concrete organizational contexts. The application of XAI in information systems is further analyzed through the lens of credibility, stability, and usefulness criteria for explanations, as well as the challenges of integrating explainability into business information systems. The paper also addresses relevant regulatory frameworks and ethical implications that drive the advancement of XAI within the domain of information systems.**

***Key Words* — Scenario, Explainable Artificial Intelligence, Information System, Transparency, Auditability, Organizational Domain, GPT-5.**

I. INTRODUCTION

In the past decade, artificial intelligence has experienced exceptionally rapid progress, both in theoretical and practical domains. A major boost to the development of artificial intelligence has come from the emergence of Large Language Models (LLMs). LLMs are based on deep neural networks, particularly on the transformer architecture. Models such as GPT-5

Filip Krstić - Author, Union University School of Computing, Serbia (e-mail: krstic.filip.raf@gmail.com).

Dušan Vujošević – Author, Union University School of Computing, Serbia (e-mail: dvujosevic@raf.rs).

(Generative Pretrained Transformer), BERT (Bidirectional Encoder Representations from Transformers), PaLM, LLaMA, and Claude have redefined the boundaries of natural language processing, enabling applications in areas that until recently were considered exclusively human — from text generation, translation, and information summarization to assistance in programming, medical diagnostics, and legal consulting. These models, trained on massive amounts of data, represent a significant leap forward in the ability of computational systems to understand and produce complex language patterns, contextual meanings, and relationships among concepts [7] [11].

However, as the power and autonomy of these systems grow, so does the need to understand their decisions, learning processes, and behaviors. In this context, the concept of Explainable Artificial Intelligence (XAI) is gaining increasing importance as a key mechanism for building trust, safety, and accountability in the use of AI systems.

As Artificial Intelligence (AI) become increasingly integrated into the informational infrastructure of society — from automated customer services and educational platforms to legal and medical assistants — concerns are also growing regarding their transparency, accountability, and ethical usability. The key question that arises is: how can we trust a system whose decision-making process we do not fully understand? It is precisely from this need that the importance of explainable artificial intelligence emerges — a field dedicated to developing methods and techniques that enable the interpretation and understanding of the behavior of complex AI models.

Traditional AI systems, such as decision trees or linear regression, sometimes allow for a somewhat simpler interpretation of their decisions. Modern models, and especially LLMs, function as highly nonlinear, multilayered systems with an enormous number of parameters operating simultaneously. Their architecture, based on the transformer mechanism, enables sophisticated contextual processing but at the same time makes it more difficult to understand the internal decision-making mechanisms.In this sense, AI models are often described as “black boxes” – systems that produce results without providing a clear insight into how those results were obtained.

The introduction of XAI approaches in working with AI systems carries multiple forms of significance. First, it enables users to build trust in the model, as understanding its decisions increases the sense of control and security. Second, it allows for the identification of potential errors, biases, or undesirable behavioral patterns within the model. Furthermore, in the regulatory context, explainability can even become a legal requirement,

particularly in cases where AI is used in so-called high-risk systems, as defined in the draft of the European Artificial Intelligence Act (AI Act) [1].

1. Explainable Artificial Intelligence

XAI represents a set of methods, techniques, and principles aimed at enabling the understanding of the processes and decisions made by complex AI systems. In the context of modern deep learning models—and especially large language models—XAI emerges as a response to the growing need for transparency, trust, and accountability in automated decision-making.

Unlike, for example, decision trees, modern models function as highly nonlinear functions with a large number of parameters, which makes them inherently opaque. In this context, XAI does not seek to simplify the model itself, but rather to provide additional information that enables users to understand how and why a particular decision was made.

XAI is particularly important in the context of information systems used in sensitive domains—such as healthcare, law, and finance—where understanding decisions is crucial for maintaining trust and protecting user rights [10] [18]. The following section discusses the key distinctions and classifications that define this field.

1.1 Classification of Explainable Artificial Intelligence Techniques

A wide range of XAI techniques has been developed in both literature and practice, and these can be classified according to several criteria. The most commonly used dimensions of classification include: 1) The way they integrate with the model – intrinsic vs. post-hoc approaches, and 2) The scope of explanation – local vs. global approaches.

Intrinsic methods involve designing the model itself to be interpretable "from within." This means that, during the model training process, attention is given to its structure and behavior to enable the generation of explanations. In the context of Large Language Models (LLMs), intrinsic methods include: Visualization of attention mechanisms (e.g., attention heatmaps), Probing techniques that examine internal representations within the model, and Architectural modifications that enable more transparent behavior. The advantage of intrinsic methods lies in the fact that explanations are built into the model itself; however, their applicability is limited to models specifically designed with this purpose in mind.

Post-hoc methods, on the other hand, are applied after the model has

already been trained. They do not alter the model's structure but instead analyze its inputs and outputs to generate explanations. The most well-known post-hoc methods include: LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), Integrated Gradients, and Counterfactual explanations. The key advantage of post-hoc methods is their flexibility – they can be applied to virtually any model. However, their accuracy and stability may be limited, particularly in the context of highly nonlinear models such as LLMs.

Another important classification refers to the scope of explanation. Local methods focus on explaining the model's behavior with respect to a single specific instance (e.g., Why was this text classified as "positive"?). These methods are useful for analyzing individual decisions but do not provide insight into the overall behavior of the model.

Global methods, by contrast, attempt to explain the general principles by which the model operates (e.g., Which words generally have the greatest impact on classification?). They are useful for understanding the model as a whole but often require simplifications and aggregations that may obscure important details.

In practice, it is often beneficial to combine local and global methods to obtain a more comprehensive view of model behavior. For example, local explanations can reveal anomalies in individual cases, while global explanations can help identify systemic patterns or biases [4].

2. THE ROLE OF EXPLAINABLE ARTIFICIAL INTELLIGENCE IN INFORMATION SYSTEMS

Information systems (IS) represent the backbone of modern organizations, enabling the collection, processing, storage, and distribution of data to support business processes and decision-making. Contemporary information systems increasingly integrate AI models as key components for automation, prediction, recommendation, and decision support.

The introduction of complex models—such as LLMs—into IS brings numerous advantages, including scalability, adaptability, and the ability to process large volumes of unstructured data. However, these systems are becoming increasingly less transparent, which heightens the risk of user distrust, regulatory obstacles, and potential ethical issues.

XAI thus could play a crucial role in preserving the integrity of information systems. In this context, XAI acts as a critical layer between the complex model and the end user, enabling the understanding of model

behavior, validation of results, and responsible management of automation. XAI makes the behavior of AI components within IS transparent, verifiable, and understandable, thus ensuring compliance with regulatory requirements while supporting users in making informed decisions [13].

2.1. Requirements for Transparency and Auditability

Transparency and auditability are fundamental requirements for any information system that employs AI components—particularly in sectors such as healthcare, finance, justice, and public administration. According to modern standards and legislation, including the EU General Data Protection Regulation (GDPR) and the Digital Services Act (DSA), users have the right to know on what basis a system has made a particular decision or recommendation [2] [5] [12].

XAI contributes to transparency by enabling the tracing of the logic behind model-generated decisions—whether through visualizations, local explanations, or language-based justifications. In this way, several key requirements are satisfied: 1) Right to explanation – users, in certain situations, must have access to the reasoning behind automatically made decisions that affect them; 2) Verifiability – oversight and audit authorities should be able to analyze the system's functioning; and 3) Accountability – companies and institutions should be able to justify and evaluate model performance.

Beyond formal regulations, there is a growing need for so-called algorithmic accountability, which implies that organizations must have both technical and procedural mechanisms to explain the results generated by their systems. This includes the ability to log decisions, record input and output data, and link explanations to specific business rules and processes. Without explainability, systems effectively stay "black boxes," resulting in a loss of control, limited ability to identify errors, and reduced capacity to improve model performance [3].

2.2. Explainable Artificial Intelligence as Decision Support within Information Systems

One of the key challenges in information systems is maintaining a balance between automation and human control. In an increasing number of organizations, AI systems do not replace decision-makers but rather support them through Decision Support Systems (DSS). Within this framework, XAI becomes an indispensable tool for interaction between models and human decision-makers.

In many cases, AI models do not make final decisions but act as recommender systems that assist humans in data analysis, pattern recognition, and the selection of optimal solutions. In such scenarios, explainability enables users to understand the rationale behind the system's recommendations, thereby improving their ability to assess their relevance, accuracy, and reliability.

For instance, in a sentiment analysis system, XAI can highlight which words most strongly contributed to classifying a text as "positive" or "negative," allowing the user to determine whether the model has correctly interpreted the context. Other examples of XAI applications in decision support are shown on the Table 1.

Field of use	Example
Financial institutions	XAI allows analysts to understand the reasoning behind a loan rejection, risk assessment, or liquidity prediction.
Healthcare systems	Physicians use XAI visualizations to understand how a model arrived at a diagnosis based on imaging data, clinical findings, or patient history.
Human Resources and Recruitment	Models filter candidates, and XAI provides explanations for each ranking to prevent discrimination.
Security systems	Security uses LLM and NLP models for threat detection, while XAI enables operators to confirm the legitimacy of such detections.

Table 1: Examples of XAI applications in decision support

In information systems based on human–machine collaboration (so-called *human-in-the-loop* architectures), XAI further enhances trust, reduces cognitive load, and accelerates decision-making. Users seek not only correct answers but also understandable reasoning, which grants the model a level of trust comparable to that placed in a human expert rather than an automated machine.

Furthermore, explainability facilitates continuous system learning, allowing users to utilize explanations to correct model misconceptions (e.g., in active learning, reinforcement learning, or interactive recommender systems). In this way, XAI not only contributes to transparency but also supports the evolution of the system itself through feedback.

In the domain of Large Language Models (LLMs), XAI has additional value by: 1) Tracking reasoning flows within generated text (e.g., *chain-of-thought* approaches); 2) Displaying key tokens and attention scores during responses, and 3) Analyzing prompt engineering techniques and their influence on model decisions. Thanks to these capabilities, XAI enables the effective integration of LLMs into information systems. Within such systems, XAI is not merely a tool for text generation but an extended interactive agent whose behavior can be analyzed, controlled, and continuously improved [16].

3. METHODOLOGICAL APPROACHES TO EXPLAINABLE ARTIFICIAL INTELLIGENCE IN THE CONTEXT OF LARGE LANGUAGE MODELS

This section analyzes the key methodological approaches to XAI in the context of LLMs. This domain requires carefully designed methodological frameworks that address the specific characteristics of these models—such as a high degree of nonlinearity, complex architectures, and an enormous number of parameters [6]. Given the complexity and nonlinearity of LLM architectures, the selection of an appropriate XAI method depends on several factors: the goal of explanation (local or global), the nature of the model (closed or open), and the type of user requirements (technical expert, end user, or regulator). Each category has its own advantages and limitations.

3.1. Post-hoc Methods in the Context of Large Language Models

Post-hoc methods are applied after the model has been trained, with the aim of providing explanations without altering the internal architecture of the model. In contrast, intrinsic methods aim to integrate explainability into the model's very structure and behavior, offering insights into the decision-making processes in real time [15].

Post-hoc methods operate under the assumption that the model functions as a "black box" whose internal parameters cannot be modified. Their purpose is to explain the model's decisions by analyzing its inputs and outputs, without changing its structure. These approaches are particularly useful when working with pretrained LLMs such as GPT or Claude, whose internal parameters are not accessible to the user.

One of the most well-known post-hoc methods for local explanations is LIME (Local Interpretable Model-agnostic Explanations). LIME generates local explanations by approximating the behavior of a complex model with a simpler, interpretable model (e.g., linear regression) in the vicinity of a

specific instance.

In the context of LLMs, LIME can identify which words in a text contribute most to a given classification (e.g., sentiment or topic). Advantages of LIME include: 1) Model-agnostic nature – it can be applied to any model; and 2) Intuitive visualization – it enables clear visualization of the contribution of individual tokens. Limitations of LIME are: 1) Instability – different runs may produce different explanations; and 2) Limited scalability – it is difficult to apply to longer texts and more complex tasks.

Another major post-hoc approach is SHAP (SHapley Additive exPlanations), which draws from game theory—specifically the concept of Shapley values—to quantify the contribution of each feature to the model's decision. In an NLP context, each word or token is treated as a "player," whose contribution is measured relative to all possible combinations of other words.

This method quantifies how much the model's prediction would change if a particular variable were removed from the input. SHAP thus enables a quantitative assessment of token importance and provides visualizations such as bar charts or heatmaps to illustrate these contributions.

A third commonly used approach is Integrated Gradients (IG), a gradient-based technique designed for neural networks. It determines which input attributes most strongly influence the model's output. Essentially, IG compares the model's prediction against a baseline value and integrates the gradients along the linear path between the baseline and the actual input. In other words, IG compute the integral of the gradients along a path from a "neutral" input to the actual one. In the context of LLMs, IG is particularly valuable because it can be applied directly to embedding layers, providing insight into the relative importance of individual tokens for the model's output.

3.2. Intrinsic Methods in the Context of Large Language Models

Unlike post-hoc methods, intrinsic approaches rely on the model's internal structure and aim to extract explanations directly from it. These methods are especially relevant for LLMs, as they exploit architectural features such as attention mechanisms and representations within hidden layers.

Intrinsic methods are designed to create models whose structure and behavior are inherently interpretable—either during or immediately after the inference process. Inference refers to the process in which a model applies previously learned patterns to generate responses, predictions, or

classifications based on new input data.

In the case of LLMs, intrinsic methods leverage internal components such as: 1) Attention mechanisms, which highlight the relationships between tokens; 2) Latent representations, which encode semantic and syntactic features; and 3) Probing models, which "interrogate" hidden layers to uncover how linguistic or conceptual information is represented and processed. These approaches allow researchers and practitioners to better understand how models represent and transform information internally—thus providing a foundation for transparency, trust, and model interpretability in real-world applications.

To analyze more deeply how a model utilizes input information during response generation, one can apply visualization of the attention mechanism from the model's final layer. The attention mechanism in transformer architectures enables the model to evaluate the importance of each word in the input sequence relative to all others, thereby forming an attention matrix that illustrates the relationships among tokens. The visualization can be represented as a heatmap, where the color intensity indicates the degree of attention the model assigned to specific words during the generation of a response.

A model's ability to identify relevant parts of the input generally confirms that the attention mechanism plays an important role in context understanding. However, attention is not always exclusively focused on key words—some portion of attention is also distributed to less relevant tokens, such as auxiliary words and punctuation marks.

This distribution of attention highlights a well-known limitation of the attention mechanism as an explanatory tool: the fact that a model "pays attention" to a given token does not necessarily imply a causal influence of that token on the model's decision. This distinction between attention and causality represents one of the key methodological challenges in XAI research. Causality has long been regarded as notoriously difficult to interpret [14].

Although attention visualizations can be useful for intuitively understanding model behavior, their interpretation must be careful and context-aware. In other words, an attention map might show, for example, that GPT-2 successfully identifies semantically relevant parts of a question, but it does not provide direct insight into how those parts are used during the generation process. The attention map indicates where the model is looking, but not necessarily why it makes certain decisions.

In the context of information systems, such visualization can serve as part

of an interactive explanation presented to users—for example, in educational applications, decision-support systems, or content analysis tools. However, its application must be carefully designed to avoid misinterpretation. Users should be clearly informed that attention does not represent a causal mechanism, but rather an indication of how the model distributes focus during input processing.

Intrinsic XAI methods such as attention visualization can significantly contribute to understanding the behavior of generative LLMs, but their interpretation requires additional analysis and contextual expertise. When combined with other techniques, attention visualizations can be a valuable component of model explainability—but they should not be used in isolation as the sole source of explanation. In information systems employing LLMs, such tools can enhance transparency and interactivity, but only if their limitations are clearly communicated and the explanations are adapted to the knowledge and needs of end users.

4. SCENARIOS FOR THE APPLICATION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE IN THE DOMAIN OF INFORMATION SYSTEMS

In this chapter, ideas for the development and testing of concrete applications of XAI will be presented. On the one hand, the scenarios refer to different organizational domains. On the other hand, they concern the comparison between alternative explainable techniques.

4.1. Scenarios of Possible Applications of Explainable Artificial Intelligence

The following scenarios outline potential directions for the development of solutions and the ways in which they could be tested. They are designed to be as precise as possible. The general application of XAI is not difficult to understand; however, concrete development requires specific, well-defined use cases that are deeply embedded in the corresponding organizational domain. Such examples must be derived from real-world practice and practical needs.

The examples refer to human resources (Table 2), the financial sector (Table 3), content delivery networks (Table 4), organizational development (Table 5), and product management (Table 6). In these examples, the need for explainability is examined in the context of different core techniques,

including both traditional and generative artificial intelligence.

Domain: Human Resources Management
Example: Explaining the orange light indicator assigned to an employee in the prediction of whether they would be a motivating leader
Possible core technique: Probabilistic classification (Naïve Bayes)
Application scenario: The prediction of whether a non-managerial employee would make a motivating leader can be performed using a model trained on data about previous and current team leaders within the company. The prediction results may be expressed as recommendations in the form of a green light icon next to the employee's name, a red-light warning, or an orange light as a mild caution. Particularly when planning to promote an employee who has received an orange light to a team leader position, it may be useful to weigh the arguments both in favor of and against the promotion.
Data in focus: length of service, productivity data, training participation records, self-evaluations, supervisor evaluations

Table 2: A human resources application of EAI scenario

Domain: Credit Risk Assessment
Example: Explaining the impact of drift on changes in the estimated probability that a client will delay payment of the next installment
Possible core technique: Logistic Regression
Application scenario: The performance of a machine learning model used to assess a client's loan repayment reliability is not static over time and may degrade gradually. Among the reasons for updating such models are slow but significant changes known as drifts. Data drift refers to changes in the distribution of input data — for example, the average income of new clients gradually increasing — while concept drift denotes a change in the relationship between variables, such as education level ceasing to be a strong indicator of risk. When a modification of the machine learning model leads to a change in the credit risk estimate, it can be valuable to check for the potential influence of such drifts.
Data in focus: client income, average income, demographic data, previous loan repayment records

Table 3: A financial sector application of EAI scenario

Domain: Multimedia Information Systems
Example: Explaining possible reasons for premium subscriber churn
Possible core technique: Clustering
Application scenario: Content delivery networks, like telecommunications providers, aim to predict which clients are likely to cancel their subscriptions. The customer relations department may prepare special offers tailored to such clients, following the principle that retaining an existing customer is easier than acquiring a new one. Understanding the reasons behind the tendency to cancel can help adapt the offer to the client's specific needs. It is particularly valuable to retain subscribers with the most expensive plans, such as premium packages.
Data in focus: daily visits, monthly visits, user activity during sessions, previous payments

Table 4: A content delivery network application of EAI scenario

Domain: Strategic Planning
Example: Explaining individual items within a generated strategic plan for a given business domain
Possible core technique: Large Language Models
Application scenario: Within the prompt interface of a chat tool providing access to a large language model, one may request the generation of a strategic plan for a specific business domain. The chatbot is provided with information such as the company's industry, its current operations, the characteristics of the business domain, and existing guidelines or ideas for further development. Generative AI incorporates these inputs into a proposed strategy, enriching them with information about current trends obtained from organizational websites, recommendations from academic literature, social media commentary, policy and regulatory guidelines, technical documentation, financial reports, and other sources. If the organization is uncertain about the validity of specific plan elements, it may request links containing evidence and arguments supporting them.
Data in focus: up-to-date web data, trend analyses from web search mechanisms, scraped meta-analytical academic papers

Table 5: An organizational development application of EAI scenario

Domain: Product Development
Example: Explaining a generated prototype visualization through brief textual annotations within the image
Possible core technique: Generative Adversarial Networks
Application scenario: Online advertising serves not only as a means of promoting a finalized product but also as an experimental tool for assessing the feasibility of developing a new one. A key aspect of online advertising is its visually compelling graphic design, which can be enhanced by illustration generation services. When developing and comparing advertising prototypes, attention can be paid not only to the overall impression but also to finer details. The evaluation process can be made faster and more effective by overlaying optional textual comments directly onto the prototype image, providing interpretative notes about specific design elements.
Data in focus: prompt data, metadata used in network training, data from large language models

Table 6: A product management application of EAI scenario

4.2. Scenarios for Comparing Explainable Techniques

For example, by using a pre-trained DistilBERT model available through the Hugging Face library, the application of explainable artificial intelligence (XAI) can be demonstrated through a task of ranking job candidates. The input text provided to the model would represent a summary of a candidate's resume, including education, work experience, and key skills. While the model might assign a high relevance score to a specific candidate, the score alone is insufficient for understanding the model's behavior—particularly in the context of information systems that require transparency and auditability.

Therefore, two post-hoc XAI methods—LIME and SHAP—could be applied to analyze which parts of the text influenced the model's decision. It is expected that LIME would identify expressions related to relevant experience and technical skills as the most influential for a positive evaluation, while SHAP could further quantify the contribution of each segment of the résumé, representing them through numerical values and visualizations. This approach would allow users of information systems to understand the criteria upon which the model bases its decisions, which is essential for building trust and fulfilling regulatory requirements in the human resources domain.

A comparative analysis of results would reveal whether there is a high degree of alignment between LIME and SHAP in identifying key tokens, indicating the consistency of explanations. It would also be interesting to examine differences in stability and granularity with respect to small textual changes. Would the removal or addition of a single word significantly alter the explanation? What would be the computational resource requirements of the two methods?

The comparison could also be extended to the differences in business user expectations. If the goal is to provide users with quick and intuitive explanations, one method may prove superior to the other, whereas if auditability is the primary objective, the situation could be reversed.

A somewhat different example would involve the use of a pre-trained GPT-5 model to explore how this large language model generates responses to prompts requesting explanations of previously produced outputs. To what extent do these explanations include references to scientific or professional literature? How well can the model reproduce reasoning processes by incorporating known mechanisms of logical inference? Which tokens in the prompt receive attention through the model's attention mechanisms? What are the explainability performance characteristics when responses rely not only on publicly available data but also on internal organizational information? How do expectations of explainability differ between general users and specialized users, such as those from a legal office? How to integrate interfaces of information systems, e.g. ERP solutions, and XAI? And finally, how many prompt iterations are required before a human user arrives at a satisfactory explanation?

5. CONCLUSION

This study explored explainable artificial intelligence (XAI) in the context of both traditional techniques and large language models (LLMs). One of the key insights highlights the need to balance the complex dynamics between the pursuit of high performance in LLMs and the growing demand for explainability. Despite technological progress, explainability remains in its infancy and is still methodologically underdeveloped.

The analysis of available XAI methods shows that the most widely adopted are post-hoc techniques such as LIME, SHAP, and Integrated Gradients. These are valuable for local understanding of model decisions; however, their application to LLMs faces serious limitations—particularly

concerning scalability, semantic depth, and interpretability for end users. On the other hand, intrinsic methods, including attention mechanisms and probing techniques, reveal potential for a deeper understanding of the internal logic of LLMs, but require a high level of expertise for interpretation and often fail to meet practical transparency requirements. The visualization and evaluation of XAI outputs pose an additional challenge, as there is still no universally accepted standard for measuring the quality of explanations.

The development of new scenarios for designing and testing XAI could focus on several domains identified in the literature as particularly promising for explainability, including healthcare [10][18], agile management and supply chain operations [17], fraud detection [9], and bioinformatics [8].

Future work should primarily advance toward the implementation and testing of solutions derived from the proposed scenarios. Solutions that reach the minimum viable product (MVP) stage should further be evaluated from a human–computer interaction perspective. In addition to ensuring user-friendly interfaces, it is equally important to avoid the illusion of transparency—a situation in which incomplete or misleading explanations conceal the model's true opacity.

Given the increasing impact of AI systems on society, explainability should no longer be treated merely as a technical add-on. It is becoming a market advantage, an ethical obligation, and a legal requirement.

Literature

(1) -, Article 6: Classification Rules for High-Risk AI Systems | EU Artificial Intelligence Act.
(2) -, European Commission. (2024). Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act).
(3) -, ISO/IEC. (2023). ISO/IEC 42001: Artificial Intelligence – Management System Standard. International Organization for Standardization.
(4) -, NIST. (2023). AI Risk Management Framework (AI RMF 1.0).
(5) -, OECD. (2019). OECD Principles on Artificial Intelligence.
(6) Bender, E. M. et al. (2021). On the Dangers of Stochastic Parrots: Can Language models be too Big? FAccT '21 Proceedings.
(7) Brown, T. B. et al. (2020). Language Models are Few-Shot Learners.

(8) Budhkar, Aishwarya, et al. "Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics." Computational and Structural Biotechnology Journal (2025).
(9) Buyuktepe, Okan, et al. "Food fraud detection using explainable artificial intelligence." Expert Systems 42.1 (2025): e13387.
(10) Černevičienė, Jurgita, and Audrius Kabašinskas. "Explainable artificial intelligence (XAI) in finance: a systematic literature review." Artificial Intelligence Review 57.8 (2024): 216.
(11) Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning.
(12) Goodman, B., & Flaxman, S. (2017). EU regulations on algorithmic decision-making and a “right to explanation”. AI Magazine
(13) Guidotti, R. et al. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys
(14) Lipton, Z. C. (2018). The Mythos of Model Interpretability. Communications of the ACM, 61(10), 36–43.
(15) Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS)
(16) Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
(17) Sadeghi, Kiarash, et al. "Explainable artificial intelligence and agile decision-making in supply chain cyber resilience." Decision Support Systems 180 (2024): 114194.
(18) Sadeghi, Zahra, et al. "A review of Explainable Artificial Intelligence in healthcare." Computers and Electrical Engineering 118 (2024): 109370.