

DOI: 10.17323/2587-814X.2026.1.7.28

Explainable AI for Industry 5.0: Shedding light on the black box

Sergey Mikhailovich Avdoshin 

E-mail: savdoshin@hse.ru

Elena Yuryevna Pesotskaya 

E-mail: epesotskaya@hse.ru

HSE University, Moscow, Russia

Abstract

The rapid development of artificial intelligence (AI) is accompanied by increasing computational complexity and decreasing model transparency, which significantly limits its adoption in critical domains that require a high level of trust, interpretability, and justification of decisions. Under these conditions, the field of Explainable Artificial Intelligence (XAI) has gained particular importance as it focuses on approaches and technologies that enable understanding of AI system logic and interpretation of their outputs. This article examines the timely topic of implementing XAI in the context of Industry 5.0. Special attention is given to practical application scenarios: the authors present concrete industrial cases from IBM, Siemens, and other companies demonstrating how XAI contributes to enhancing the reliability, safety, efficiency, and trustworthiness of AI systems. The study includes a systematic search and analysis of the literature in this domain and proposes well-grounded key criteria for comparing existing XAI approaches. The article also outlines the advantages, current limitations, and promising directions for the development of XAI, highlighting the opportunities it opens for improving effectiveness, transparency, and trust in business.

Keywords: XAI, explainable artificial intelligence, Industry 5.0, machine learning, industry

Citation: Avdoshin, S. M., & Pesotskaya, E. Yu. (2026). Explainable AI for Industry 5.0: Shedding light on the black box. *Business Informatics*, 20(1), 7–28. <https://doi.org/10.17323/2587-814X.2026.1.7.28>

Introduction

The emergence of explainable artificial intelligence (XAI) is directly associated with the rapid progress of modern machine learning methods, particularly deep neural networks. These models have demonstrated outstanding performance across a wide range of tasks; however, they have also come to be perceived as so-called “black boxes”, that is, highly complex systems whose internal mechanisms are largely opaque to users [1]. In contrast to earlier AI systems, such as expert systems or rule-based models that were relatively transparent, contemporary deep learning algorithms contain millions of parameters. As their complexity has increased, interpreting the decisions they produce has become nearly impossible in practice. This has given rise to what is often described as an “explainability barrier”, which limits the adoption of AI due to insufficient trust in opaque models [2].

Modern society expects artificial intelligence to be not only effective, but also reliable, transparent, and fair [3–5]. A lack of clear explanations for algorithmic decisions leads to concerns among users as well as regulatory authorities.

Explainable artificial intelligence has emerged as a response to this challenge. Its primary aim is to improve the interpretability and transparency of AI black-box models. XAI seeks to bridge the gap between the growing complexity of modern algorithms and the human need to understand the results they generate. Within the XAI paradigm, methods, techniques, and algorithms are developed to provide interpretable and intuitively meaningful explanations of AI-driven decisions. In this way, XAI offers developers, users, and regulators clear and well-reasoned explanations.

Within the human-centric vision of Industry 5.0, XAI is regarded as a key enabler of successful AI deployment. It allows users to understand and trust algorithmic outcomes, which is essential for effective human-machine interaction. Explainable AI also helps ensure that digital systems remain ethical, accountable, and aligned with human values and objectives [6, 7].

For business leaders, XAI is no longer merely a technical add-on, but a necessary condition for effective decision-making and governance. As algorithmic complexity increases, black-box models deprive managers of the ability to assess the rationale behind decisions that underpin strategic and operational actions. The adoption of XAI helps address this challenge by providing transparent explanations of algorithmic behavior. This supports more informed and responsible decision-making, reduces organizational risks, and creates new opportunities for innovation and development. For companies seeking to remain competitive in the context of Industry 5.0, the implementation of XAI becomes a strategic necessity [8–10].

In this study, the authors analyze contemporary approaches and requirements related to explainability that aim to enhance the transparency and reliability of intelligent systems and to strengthen trust in their decisions. A systematic literature review on XAI was conducted based on defined inclusion and exclusion criteria, analysis of citation databases, and structured synthesis of the selected publications. Section 1 examines the nature of XAI in the context of Industry 5.0, discusses its role and the black-box problem in business applications, and compares existing approaches. Section 2 focuses on opportunities for applying XAI in business and key directions for its adoption. Section 3 presents practical cases and

industry examples demonstrating the effectiveness of XAI in corporate settings. Section 4 analyzes barriers and limitations that hinder the widespread adoption of XAI and assesses associated risks. Finally, Section 5 discusses promising directions for future development and potential trajectories for the use of XAI in business decision-making.

1. The Concept of explainable AI in the context of Industry 5.0

1.1. Industry 5.0 and explainable AI

The widespread adoption of artificial intelligence (AI) in critical domains has revealed a number of challenges related to explainability, particularly in the context of Industry 5.0. The European Commission defines Industry 5.0 as a model of industry that complements the existing Industry 4.0 paradigm with a human-centric approach and resilience to external disruptions [11]. While Industry 4.0 primarily focused on technologies such as autonomy, digital connectivity, and data-driven processes, Industry 5.0 places humans at the center, emphasizes close integration with AI, and incorporates social responsibility as a core principle. Industry 5.0 positions human involvement as a key element of production and management processes [11, 12] and promotes closer collaboration between humans and AI or robotic systems in the workplace. In this paradigm, humans are not removed from decision-making processes; instead, technologies are designed to augment human capabilities, enhance comfort and safety, and enable personalized production tailored to individual needs.

Under these conditions, XAI becomes a crucial factor for both trust and effectiveness, serving as a bridge between the growing complexity of modern black-box models and the demand for reliable and transparent AI systems. XAI is commonly defined as the ability of a system to provide human-understandable explanations of how decisions are made [13]. Its goal is to make AI models transparent, interpretable, and trustworthy by explaining both the internal processes and the outputs of algorithms [14].

The motivation for developing XAI in business is largely driven by ethical and legal considerations. First, regulators increasingly impose requirements for algorithmic transparency. In the European Union, the concept of a “right to explanation” for decisions made by automated systems is actively discussed. For example, in the banking sector, if a loan application is rejected by an automated decision-making system, the client has the right to be informed about the reasons behind that decision [15]. Such regulations, including the requirements of the General Data Protection Regulation (GDPR), compel organizations to implement explainability mechanisms; otherwise, the use of black-box models may entail legal risks and consequences [16].

Second, socio-organizational factors also play a significant role. As noted by Zavodna et al. [17], insufficient transparency of AI systems leads to resistance among users and managers during implementation. In business practice, there is growing evidence that opaque AI systems are often rejected by organizations, ultimately reducing the effectiveness of digital transformation initiatives.

Ensuring explainability is therefore a necessary condition for building trust in AI among employees, customers, and service users. According to recent studies [18, 19], XAI helps identify and mitigate model biases, ensure compliance with ethical standards, and improve the justification of algorithmic decisions. As a result, explainability increases users’ willingness to accept and effectively utilize AI-based systems. It can be concluded that within the human-oriented paradigm of Industry 5.0, where machines are intended to complement rather than replace humans, transparency of AI decisions becomes a prerequisite for safe and productive human-AI collaboration.

As part of this study, a systematic literature search and analysis on XAI was conducted based on defined inclusion and exclusion criteria, citation database analysis, and structured organization of the selected materials (*Fig. 1*). The research is grounded in a comprehensive review and analysis of scientific literature on explainable artificial intelligence and its applications in business and Industry 5.0.

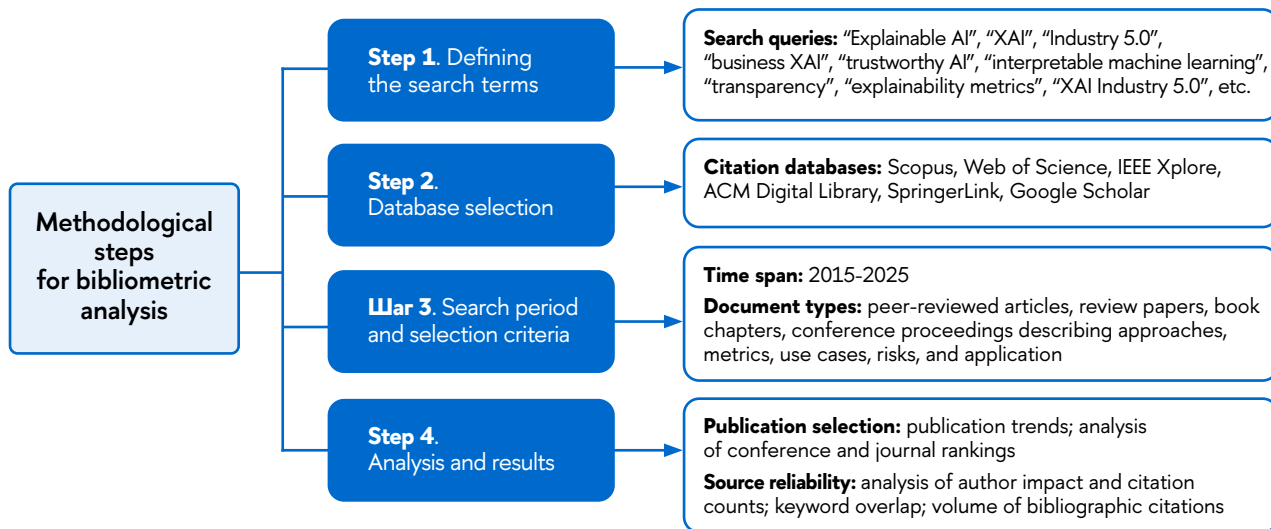


Fig. 1. Bibliometric analysis methodology.

The most significant growth in publications is observed in the period from 2021 to 2024, which can be attributed to the increasing business interest in algorithmic transparency and evolving regulatory requirements.

The thematic structuring was carried out along the following directions, which form the basis of the structure of this study:

- ◆ conceptual foundations of XAI;
- ◆ methods and metrics;
- ◆ applications in business;
- ◆ barriers and risks;
- ◆ regulatory aspects.

It is important to note that explainability is not a single or static attribute. In academic research on XAI, it is treated as a complex, multidimensional criterion that encompasses a range of aspects, from model transparency (the extent to which its internal mechanisms are accessible for understanding) and interpretability (the extent to which one can understand why a specific decision was made), to accuracy, fairness, the faithfulness of explanations (i.e., avoiding misleading rationales), and accountability. For example, a simple and

transparent model may be easy to understand, but not necessarily accurate. For this reason, each XAI project must strike a balance between these dimensions.

Joyce et al. [20] propose to view explainability as a function of comprehensibility that reflects both transparency and interpretability. Arrieta et al. [14], along with Murdoch [21], place these concepts within a broader framework of responsible AI, extending them with notions such as trust, reliability, and related considerations. In this article, the authors propose an original map of the most commonly discussed explainability properties (Fig. 2).

At present, there is no single, widely accepted standard that defines what should be considered explainability in artificial intelligence. This multidimensionality reflects the inherent complexity of the concept itself and highlights the need for systematization and alignment of terminology and evaluation approaches for XAI, depending on the application context, including the industry, model type, and target audience.

There is also no universal set of quantitative or qualitative metrics for measuring the level of explainability. Different approaches rely on different criteria, ranging from subjective user understanding of explana-

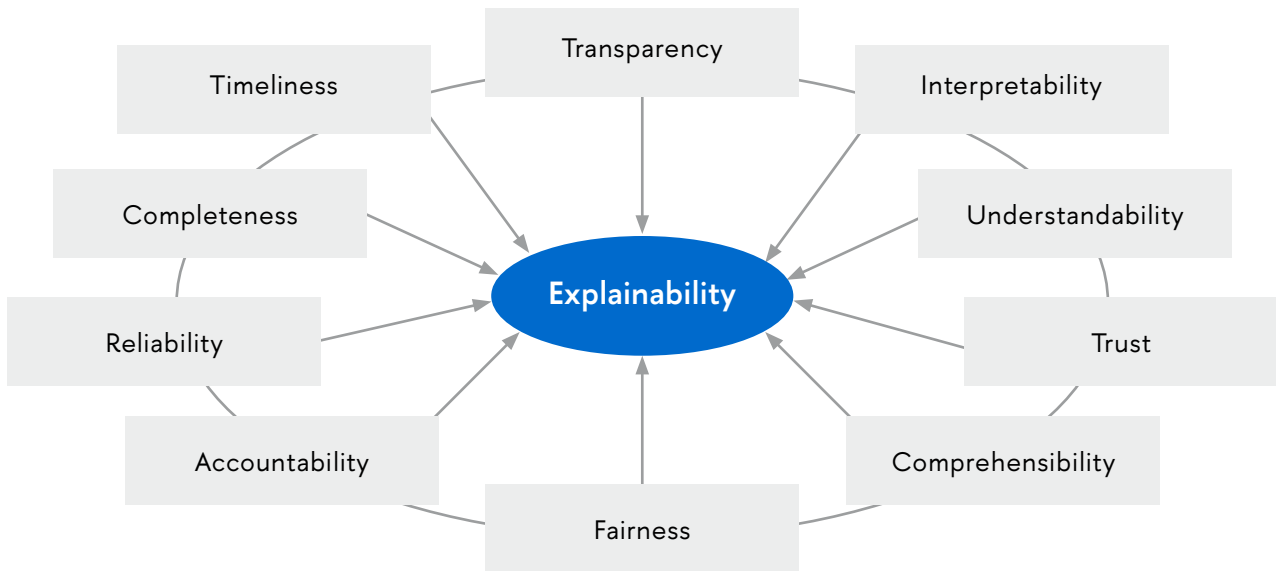


Fig. 2. Overview of key explainability characteristics.

tions to formal measures of stability and local fidelity of interpretations [22, 23]. As a result, the adoption of explainable AI requires not only technical implementation, but also methodological effort to determine what constitutes “sufficient” explainability in a given context.

Despite the diversity of existing approaches, it can be argued that the comparison of explainability methods is based on a number of recurring criteria [13, 14, 20, 21]. The following characteristics are most frequently discussed in the literature:

- ◆ Type of explanation (local vs. global, post-hoc vs. built-in / intrinsic).
- ◆ Transparency and interpretability (the extent to which a human can understand the model’s reasoning).
- ◆ Stability of explanations (the degree to which explanations change in response to small variations in input data).
- ◆ Accuracy and informativeness of explanations (whether the explanation correctly reflects the model’s underlying logic).
- ◆ Robustness to noise and adversarial attacks (high or low robustness).

- ◆ Temporal and computational complexity, which is particularly important for industrial applications and real-time processing (high, medium, or low).

These criteria are widely discussed in international studies and are commonly used as a basis for the systematic comparison of XAI approaches.

1.2. The “black box” problem and AI trust in business

The demand for responsible, transparent, and human-centric artificial intelligence in business is steadily growing, in line with current technological and managerial trends. Best practices are gradually emerging, industry-specific guidelines are being developed, primarily in the financial sector and healthcare, and many companies begin to define explainability requirements independently, based on the specifics of their business processes.

In managerial and business domains, the use of artificial intelligence is already becoming routine, ranging from recommendation systems in e-commerce to algo-

rithms for business process optimization and data-driven decision-making. However, the willingness of companies and organizations to entrust critical processes to black-box models remains limited.

A black-box model refers to a system whose internal structure is hidden or too complex to be understood by humans. Many high-performing machine learning algorithms, including deep neural networks and ensemble methods, are characterized by a high degree of opacity. While such models were previously applied to narrowly scoped tasks, this lack of transparency was not considered critical. However, as AI systems have increasingly been deployed in domains that directly affect human decision-making and activities, opacity has become a fundamental issue, as emphasized in a number of XAI studies [24–26]. As AI systems began to influence decisions with real-world consequences, the lack of transparency turned into a serious concern. For businesses, a key question arises: if the internal workings of an algorithm are not understood, can its decisions be trusted?

The absence of explanations makes it difficult to detect errors and biases. Hidden biases in training data may lead to unfair or discriminatory outcomes. For example, algorithms used for recruitment or customer selection may unintentionally discriminate against certain groups of applicants.

When it comes to trust in artificial intelligence, especially in high-risk domains such as transportation, finance, and industry, people tend to reject even highly accurate model outputs if no rational explanation is provided. In the banking sector, for instance, both clients and managers expect to understand the reasons behind a loan rejection. In industrial settings, the absence of explanations may result not only in distrust, but also in critical safety risks if a model fails. Any decision of significant importance must therefore be accompanied by explanations that are understandable to humans. Without this, the use of black-box models in such domains is generally considered unacceptable.

A trade-off between model accuracy and interpretability is widely discussed in the literature. Indeed,

the most transparent models, such as shallow decision trees, often demonstrate lower accuracy on complex tasks compared to deep neural networks. Conversely, efforts to maximize performance metrics frequently result in highly complex models at the expense of interpretability. In practice, this leads either to a simpler, so-called ‘gray-box’ models with limited performance, or to powerful black-box models whose high accuracy is achieved through a loss of transparency [27, 28]. The objective of XAI is to reduce this gap by offering methods that preserve predictive performance while providing meaningful explanations. However, this trade-off has not yet been fully resolved, and the question of how much performance can be sacrificed for the sake of transparency remains open.

In summary, the black-box problem is not merely a technical metaphor, but a serious barrier to the widespread adoption of artificial intelligence across business domains.

2. Business needs for explainable artificial intelligence

Based on the results of the bibliometric analysis, the authors found that the most densely represented application domains of explainable artificial intelligence today are finance and industry. In these areas, XAI is used both to build trust in algorithmic systems and to support operational and strategic decision-making. The energy sector, public administration, and healthcare appear predominantly in studies of a regulatory and normative nature, reflecting the growing attention to transparency, auditability, and non-discrimination of algorithms in high-risk domains.

Logistics is emerging as a new and rapidly developing area of XAI application. Although the number of publications in this field remains relatively limited, the domain demonstrates strong growth and increasing interest from both researchers and companies in explainability for supply chain systems and intelligent optimization (*Table 1*).

Table 1.

Bibliometric analysis of XAI application areas in various industries

Source	Finance	Industry	Energy	Business	Economics	Management	Logistics	Public sector	Healthcare	Research focus and key ideas
1. Martins et al. (2024) [10]	X			X	X	X				Overview of XAI in finance; SHAP/LIME; transparency of credit scoring
2. Gramegna & Scardapane (2021) [31]	X					X				Discrimination assessment; explainability in credit risk
3. Hjelkrem & de Lange (2023) [32]	X					X				Explaining deep learning models in open banking
4. Poyiadzi et al. (FACE, 2020) [30]	X	X		X	X		X			Counterfactual explanations; applicability across domains
5. Weitz et al. (2022) [18]				X		X				AI acceptability; reduction of organizational resistance
6. Chehbi-Gamoura (2023) [6]				X		X				Explainability in decision-making
7. Tabassi (NIST AI RMF, 2023) [16]	X	X	X	X	X	X		X	X	Regulatory requirements for XAI
8. EU AI Act (2024) [41]	X	X	X	X	X	X		X	X	Legal requirements for explainability
9. Ahmed et al. (2022) [12]		X		X	X		X			XAI in Industry 4.0 / Industry 5.0
10. Adadi & Berrada (2018) [13]	X	X		X	X	X			X	XAI taxonomy; post-hoc methods
11. Arrieta et al. (2020) [14]	X	X	X	X	X	X		X	X	Responsible AI; properties of explainability
12. Černevičienė & Kabasinskas (2024) [42]	X			X	X	X				Systematic review of XAI in finance; tasks: scoring; SHAP/ANN/XGBoost methods
13. Brasse et al. (2023) [43]				X	X	X				XAI in information systems; classification of research directions
14. Samek W., Montavon G., et al. (2019) [1]	X			X	X	X				Survey of XAI methods; taxonomy of rule-based, model-agnostic and intrinsic models; widely cited
15. Carvalho et al. (2019) [44]		X		X	X		X			Systematic review of XAI challenges; industry-focused limitations and opportunities
16. Molnar (2025) [45]	X	X		X		X				Comprehensive interpretability; model-agnostic techniques; stability of explanations
17. Angelov et al. (2021) [46]	X	X		X						Explainable-by-design methods; self-explainable models; interpretable fuzzy-rule
18. Rai (2020) [47]	X			X	X	X				Explainable AI in management and decision support; accountability frameworks
19. Samek & Müller (2017) [22]		X	X	X						Visualization techniques; explainability evaluation; relevance propagation
20. Liao & Varshney (2021) [48]	X	X		X		X				Human-centered XAI; user modeling; adaptive explanations for stakeholders
21. Chamola V et al. (2023) [49]		X	X	X	X	X	X			Explainability in cyber-physical systems; domain-aware explanations; OT/IT integration
22. Belle & Papantonis (2021) [50]				X	X	X				Logic-based XAI; symbolic reasoning; foundations for transparent decision-making

The integration of explainable algorithms into real-world practice helps mitigate the black-box problem and makes AI systems more understandable and acceptable for business use. Requirements for explainable AI in business, economics, and management can be classified into several key dimensions that reflect both the practical needs of organizations for explainable algorithms and the demands imposed by the external environment.

A. Trust and transparency of decisions. Trust is regarded as a fundamental prerequisite for the functioning of artificial intelligence systems in the digital economy [29]. Explainable AI can provide explanations in a human-understandable form, thereby reducing the risks of mistrust and discrimination, and justifying AI-based recommendations to clients, shareholders, auditors, and employees. For example, XAI-based credit scoring systems can detail the contribution of individual factors such as income level or credit history, thus meeting regulatory requirements and strengthening customer trust. In human resource management, explainability helps prevent unjustified decisions. If an algorithm filters out job candidates, a company must ensure that this occurs for relevant reasons rather than due to hidden discrimination. By providing HR specialists with interpretable criteria and information about which skills or competencies were decisive, XAI makes the selection process more transparent and fair. This reduces the risk of bias and increases employee trust in such systems.

B. Integration of AI into operational workflows. Explainability facilitates the integration of algorithms into everyday work processes, reduces staff resistance, and helps establish a shared “language” between humans and AI systems. At the organizational level, researchers introduce the concept of AI acceptability, which reflects the willingness of employees and managers to adopt and use AI tools. Empirical studies show that the main barriers are socio-organizational factors, largely related to trust and understanding [16, 43]. Employees may resist algorithmic decision-making due to fears of losing control or skepticism toward “machine-generated” outcomes. However, when systems provide clear explanations and involve users in

the decision-making process, the likelihood of mutual and trust-based collaboration increases. For instance, engineers are more likely to rely on a predictive system if it indicates which specific sensor readings led to a given forecast and provides relevant contextual information.

C. Strategic management and business analytics. XAI supports top management and business owners in strategic decision-making. Businesses increasingly rely on analytical models for strategic planning, risk assessment, and analysis of consumer behavior. However, executives are often unwilling to base decisions on model outputs if the underlying assumptions and reasoning are unclear. As a result, explainable models, such as econometric models with interpretable coefficients or advanced machine learning models enriched with XAI explanations, are preferred in corporate analytics.

Recent surveys, including the work by Tchuente, Lonlac, and Kamsu-Foguem [9], propose a structured evaluation approach based on theoretical foundations, application context, data and task characteristics, and solution methodology (TCCM: Theory, Context, Characteristics, Methodology). In real-world settings, it is important to explain the entire managerial decision-making process: why a particular question is posed, why specific data are used, how the model arrives at its conclusions, and whether domain experts validate these explanations in practice. Without human validation of explanations, the application of XAI in business remains incomplete. This is why experts recommend establishing an iterative cycle consisting of model development, explanation generation, expert evaluation of explanations, and subsequent adjustment of the model or its application.

D. Control of complex industrial systems. Modern industrial environments generate vast amounts of data, ranging from equipment sensor readings to financial and logistics information. AI models are capable of identifying hidden patterns in these data and optimizing operations. Nevertheless, engineers and operators must understand these patterns, especially when systems propose non-standard actions, such as shutting

down a machine due to a detected anomaly. XAI enables the integration of explanatory modules into industrial analytics systems, clarifying which sensors or indicators exceeded normal thresholds, why a failure is predicted, or which factor was decisive in identifying a product defect.

For these purposes, the FACE (Feasible and Actionable Counterfactual Explanations) approach has proven effective and is well documented in the literature [30]. FACE identifies realistic and achievable pathways from the current state to a desired outcome while accounting for feasibility constraints such as technological tolerances, safety requirements, and operational regulations. As a result, personnel receive explanations from AI systems in an understandable form, whether as charts, textual descriptions, or visual highlights of problematic components within system diagrams. These explanations indicate which factors were critical to the system's conclusions and what changes are required. If a robot or automated production line behaves unpredictably, this poses risks to both personnel and production safety. The availability of explanations, for example, "the robot reduced speed because a sensor detected a deviation in raw material quality," allows engineers to analyze the data and rules underlying the decision and to adjust the algorithm to prevent similar errors in the future.

E. Compliance with regulatory requirements. Many sectors of the economy are subject to strict regulation, including finance, industry, and energy. To avoid reputational and legal risks, businesses require ethical oversight mechanisms and algorithmic audit procedures. XAI tools provide technical support for these initiatives. In effect, explainability becomes a competitive advantage: companies that can demonstrate the transparency and fairness of their algorithms gain greater trust from both consumers and regulators [10], including in the context of regulations such as the GDPR [15] and the AI Act [41].

In summary, in business, economics, and management, explainable AI enhances the transparency of business analytics, improves human-algorithm interaction within organizations, and supports compliance

with ethical and regulatory standards. Explainability is gradually becoming part of corporate data culture. Managerial decisions are now expected to be not only data-driven, but also explanation-driven, that is, accompanied by clear and understandable justifications. Only when algorithmic decisions are supported by meaningful explanations are all stakeholders willing to accept and endorse them.

3. Practical applications of XAI: Use cases and industries

Explainable artificial intelligence is most in demand in domains where automated decisions have a direct impact on people, their health, well-being, rights, and safety. In such contexts, improving predictive accuracy alone is insufficient. It becomes essential to ensure that decisions are understandable and well justified, which makes XAI a critical component of AI deployment.

Below, we outline key business domains in which XAI is already being applied or actively introduced, along with representative use cases and tasks where explainability plays a decisive role.

3.1. Financial sector

Finance can be regarded as one of the core sectors and among the most heavily regulated areas of AI application. Here, explainability directly affects not only customer trust but also compliance with mandatory legal and ethical requirements. At the same time, a well-known tension exists between accuracy and interpretability: deep learning models often demonstrate high predictive performance but remain difficult to explain. To improve transparency, banks tend to rely either on more interpretable models, such as gradient boosting methods where feature importance can be assessed, or on the application of XAI techniques. These include interpretable scoring cards and monotonic Gradient Boosting Machines (GBMs), which preserve logical relationships between input factors and the final score.

Methods such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive Explanations) are gaining increasing popularity, particularly in credit scoring, investment analysis, and risk assessment, as reported in recent studies [31–33].

Credit scoring and loan approval. When credit decisions are made automatically, banks are often required to provide borrowers with reasons for rejection. Customers have the right to understand why their application was denied, while banks must ensure that model decisions are not based on discriminatory attributes such as gender, age, or ethnicity. SHAP, for instance, provides numerical estimates of feature contributions. These values can be simplified and communicated as human-readable justifications, for example: insufficient income (–20 points), short credit history (–15), high current debt (–10). The model can then analyze what changes in the input data would lead to a different outcome and suggest ways for the customer to improve creditworthiness. This approach satisfies regulatory requirements while also improving transparency and customer understanding.

Investments and trading. In investment analysis, explainability acts as a trust-building factor between the system and its users. Algorithms that generate investment recommendations must justify them in order to persuade investors to follow such advice. Investors relying on AI-based recommendations need to understand which macroeconomic or market signals underlie a forecast. Explanations may take the form of narrative reasoning, for example: “We recommend reducing equity exposure due to emerging risk signals, such as rising inflation and declining corporate earnings.” Such explanations help justify decisions and reduce both regulatory and reputational risks.

Risk analysis and fraud detection. The growing complexity of financial transactions and the constant evolution of fraud schemes require explainable solutions. In this area, XAI serves as a tool for expert validation of model behavior. Explainability helps clarify why a transaction was flagged as suspicious, for example due to an unusual geographic location or an exceeded

transaction limit. This makes it possible to distinguish genuine threats from false positives and reduces operational costs. The use of XAI in insurance analytics and risk management further improves the interaction between algorithmic outputs and expert judgment, enabling model correction, retraining, and the creation of additional competitive advantages [34].

3.2. Industry

Another key domain for XAI adoption is industry and so-called smart manufacturing, where AI is used to predict equipment failures, optimize product quality, and manage supply chains. In the context of Industry 5.0, it becomes critical not only to predict events but also to explain the reasoning behind algorithmic recommendations. This allows engineers and operators to trust system outputs and act upon them.

Predictive maintenance. Traditional maintenance approaches face multiple challenges. Algorithms often generate numerous false alarms without explaining their origin, leading to unnecessary inspections and downtime. Moreover, sensor-based data are highly dynamic: after repairs or upgrades, equipment behavior changes, reducing predictive accuracy and causing data drift. Operators may also receive opaque alerts without understanding which parameters triggered them or what actions should be taken.

Research by Watanabe et al. shows that constrained generalized additive models (GA2M+) combine strong predictive performance with a structure that is more interpretable for engineers and aligned with the physical logic of processes [35]. Risk attribution techniques for time series analysis enable the assessment of individual factor contributions within specific observation windows. Surrogate rule-based decision trees built on top of black-box models translate complex predictions into simple, operator-friendly explanations, while counterfactual explanations indicate which parameter changes would reduce failure probability to an acceptable level.

A practical example is IBM Maximo Predict [36], which uses AI together with sensor data, mainte-

nance reports, and failure histories to forecast equipment breakdowns and provide interpretable explanations to specialists. As shown by Hermans et al. [37], incorporating SHAP analysis and interpretable models into predictive maintenance systems can reduce false-positive alerts by more than 90 percent, significantly increasing engineer trust and operational efficiency.

Quality control. XAI is also becoming increasingly important in product quality inspection. Computer vision algorithms are widely used to detect defects on production lines, but traditional models often limit output to binary classifications without explaining the reasons behind them. This undermines operator trust and complicates root cause analysis. Interpretation techniques such as LIME and SHAP allow visualization of image regions that were decisive for classification. As a result, engineers can better understand system decisions, identify defect sources more quickly, and more confidently adopt automated quality control.

Logistics. Logistics and supply chain management benefit from AI in route optimization, resource allocation, and warehouse management. However, inventory and delivery optimization algorithms are often perceived by managers as black boxes, which reduces their willingness to adopt recommended strategies. Explainability helps overcome this barrier. Systems that clearly demonstrate which factors, such as demand growth, supplier delays, or transportation cost changes, influenced a decision inspire greater trust and improve alignment between human judgment and algorithmic recommendations. Experimental studies show that the use of SHAP and LIME increases transparency and trust in AI-driven logistics and inventory management decisions [38].

The concept of industrial XAI for manufacturing processes is actively promoted by Siemens. In its technical report [39], the company emphasizes explainability as a core requirement for industrial AI, stating that it must be ensured throughout the entire lifecycle of AI systems, from problem formulation to monitoring and operational support. This

example highlights the growing role of XAI as a mandatory element for transparency and governance in Industry 5.0 systems.

The range of domains in which AI can be applied is much broader. Beyond business, transparency and explainable AI are increasingly relevant in social domains such as healthcare, politics, law, and public administration, where the cost of decisions is particularly high and public trust is critical. At the same time, many areas remain less explored from an XAI perspective, especially where AI adoption is still limited. These include agriculture, such as crop yield forecasting and machinery management, energy systems optimization, the entertainment industry, where understanding audience preferences is essential, as well as culture and the arts.

4. Barriers to XAI adoption: Economic, technical, and organizational factors

Despite the clear benefits of XAI in terms of trust enhancement and risk reduction, its widespread adoption in business still faces significant barriers. A central question in the business environment is return on investment. If the benefits are not immediately visible or do not directly translate into profit growth, XAI may be perceived as an optional feature. To convince management or investors, tangible effects must be demonstrated: increased customer loyalty, reduced error rates, or lower compliance costs. Such conclusions require empirical evidence, yet compared to traditional AI deployments, there are still relatively few published cases that quantify the impact of XAI.

To further clarify and empirically validate the key barriers to XAI adoption in business and industrial contexts, a targeted bibliographic analysis of scientific publications was conducted (*Table 2*). The analysis included only studies in which XAI is examined not as an abstract technical concept, but in the context of real-world applications in organizations, industry, digital manufacturing, corporate governance, the financial sector, or the regulation of high-risk AI systems.

Table 2.

Taxonomy of XAI adoption barriers

Type of barrier based on bibliographic analysis	Source
1. Technical limitations (integration complexity, lack of standards, low performance of XAI methods)	[7], [8], [9], [10], [12], [25], [27], [28], [36], [39], [40], [43], [49], [58]
2. Organizational challenges (lack of competencies, need for staff training, process changes)	[6], [7], [8], [9], [11], [12], [17], [18], [19], [29], [36], [39], [47], [48], [43], [54], [55]
3. Economic barriers (implementation costs, return on investment (ROI), resource constraints)	[7], [8], [9], [12], [17], [18], [19], [29], [34], [36], [39], [42]
4. Regulatory and compliance barriers	[3], [4], [10], [11], [12], [15], [16], [29], [34], [36], [39], [41], [42]
5. User-related / human factors (trust, cognitive load, non-intuitive explanations)	[6], [8], [9], [11], [17], [18], [19], [29], [36], [39], [47], [48], [49], [54], [55]

The most frequently cited barrier in the literature relates to organizational challenges. The human factor, rooted in long-standing reliance on personal expertise and intuition, often leads to organizational resistance to the adoption of XAI. Not all organizations are ready to accept “advice from a machine.” Concerns arise as to whether employees will trust insights generated by AI without an adequate level of explainability. In addition, effective use of XAI requires new roles, ranging from explanation designers to interpretation specialists. Investments in staff training are therefore necessary so that employees perceive XAI as a supportive tool rather than a threat to their professional position.

Resistance may also result in slower decision-making processes. Explainable models require time for review and interpretation, which contrasts with the business drive for speed and efficiency. If not properly integrated into workflows, XAI can reduce operational agility. This creates a need for time-saving solutions, for example, reducing the number of meetings because all participants immediately understand the algorithm’s logic and spend less time debating the validity or transparency of its outcomes.

Another important barrier is the lack of personalization of explanations. Most current XAI systems provide standardized explanations without accounting for the user’s level of expertise, task, or situational context. As a result, explanations may be overly complex for some users and overly simplistic for others. Research has begun to explore adaptive explanation approaches, in which the system assesses whether the user has understood previous explanations and adjusts the level of detail accordingly. However, such methods have not yet been widely adopted in practice.

Technical limitations and resource constraints represent another major obstacle. Many XAI methods remain at the level of research prototypes, often implemented as scripts or notebooks that are difficult to integrate into production environments. These methods can be computationally intensive, slow, and dependent on access to the internal structure of models. For example, generating a single explanation using LIME may require hundreds or even thousands of model evaluations, resulting in substantial computational overhead [40]. In real-world settings, engineering teams are forced to seek compromises, such as caching, approximation techniques, or optimization of XAI

components, to ensure that explanations can be delivered in real time or within acceptable latency. Users are unlikely to tolerate long delays while a system “thinks” about an explanation.

In addition, there is a lack of widely accepted industry standards for explanation formats and no unified platform that supports all models out of the box. As a result, companies often develop XAI solutions tailored to their specific needs. This leads to duplicated efforts, increased costs, and limited scalability, as each organization invests its own time and resources into bespoke implementations.

Finally, high implementation costs and complexity pose a significant constraint. Explainability is often treated as an additional module that requires interface redesign, business process adaptation, and substantial team preparation. In financial institutions, for example, it is not sufficient to generate explanations for scoring models; employees must be trained, client-facing interfaces updated, and explanations presented in a clear and compliant manner. These associated costs can hinder adoption, particularly when XAI is not explicitly mandated by regulators or industry standards. Consequently, even where the importance of explainability is acknowledged, XAI may still be perceived as a secondary feature rather than a necessary strategic investment.

The identified barriers can be represented in the form of a risk map, allowing for the assessment of their

likelihood and potential negative impact. In *Fig. 3*, the authors highlight key risks with medium to high probability of occurrence and the most significant economic, technical, and organizational consequences associated with the implementation and use of XAI in business contexts. *Table 3* presents the mitigating measures.

To overcome the identified risks and barriers, it is necessary not only to advance technologies and system architectures, but also to establish new standards, develop relevant competencies, and adapt business processes. In addition, specialized artifacts need to be designed, which are discussed in the following sections.

5. The future of XAI adoption in business AI solutions

Industry 5.0 effectively positions explainable AI as a standard rather than an optional feature. This requirement creates the foundation for genuine human-AI partnership, which is frequently highlighted as a core principle of Industry 5.0. In the factories and organizations of the future, highly skilled operators and managers will make decisions jointly with AI systems: the AI will identify risks or optimization opportunities, explain the logic behind its recommendations, while the human expert, having understood this logic, will approve or reject the proposed action, introducing human judgment, creativity, intuition, and responsi-

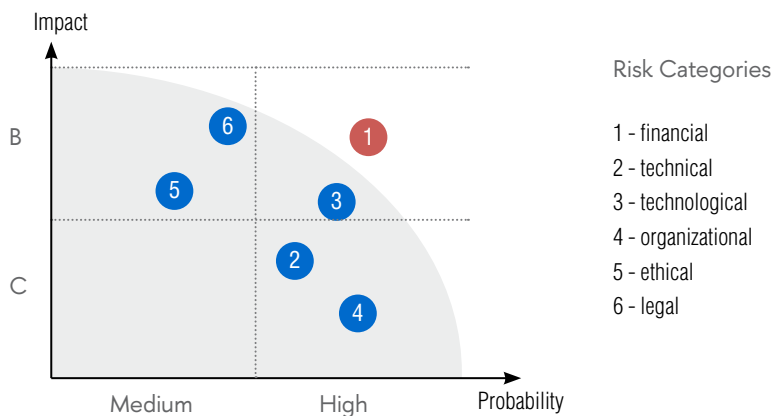


Fig. 3. Mapping of key risks for XAI adoption.

Table 3.

Mitigations	
Key Risks	Mitigations
1. Unclear return on investment, budget constraints, additional costs for XAI integration	XAI dashboards and monitoring panels; explainability KPIs; phased pilots; stop-criteria
2. Infrastructure upgrades and high computational cost of XAI methods	Approximations and caching; precomputation; hybrid models: fast rule-based components with asynchronous explanations
3. Complex integration of XAI outputs into system architecture; lack of unified standards	XAI explanation interfaces with REST APIs and JSON outputs; modular architecture solutions
4. Resistance to change, lack of XAI expertise, complex user interfaces	Training materials (XAI guidelines, workshops); two-level explanations; UX/UI protocols
5. Risks of discrimination, bias, manipulative or incomplete explanations, erosion of trust	Multi-level bias audits; removal or restriction of sensitive attributes; counterfactual testing; regular drift monitoring
6. Leakage of confidential information through explanations; lack of accountability	Control of explanation granularity and access; built-in output filtering policies; logging and audit trails

bility. Such synergy is only possible when supported by a robust XAI infrastructure combined with appropriate organizational changes.

5.1 Development of regulatory requirements

Regulatory pressure remains one of the most powerful drivers of XAI adoption in business. In Europe, the proposed AI Act places a strong emphasis on explainability, transparency, and auditability, defining them as fundamental requirements for high-risk AI systems, ranging from financial decision models to industrial applications. Legislators are gradually formalizing expectations regarding algorithmic explainability, particularly in critical domains.

In the financial sector, for example, regulations may require that all decision-making models provide clients with clear information about the key factors influencing outcomes. Future scoring and trading systems will be expected to demonstrate transparency, avoid discriminatory behavior, and explicitly disclose associ-

ated risks. A promising direction is the emergence of self-diagnostic algorithms that not only explain their outputs but also automatically check for prohibited dependencies and integrate explanation generation directly into reporting and compliance workflows.

5.2. Human-machine symbiosis

Modern XAI systems must be context-aware and capable of adapting explanations to domain-specific logic. This requires close collaboration with subject-matter experts and the incorporation of domain knowledge, including ontologies and formal rules.

Seminal works by d’Avila Garcez et al. and Besold et al. propose approaches for embedding logical reasoning into neural networks through neuro-symbolic AI methods, which combine deductive reasoning with deep learning. These approaches enable models to rely on explicit rules, ontologies, and domain knowledge, thereby improving interpretability, traceability, and documentation of reasoning processes.

In the future, business decision support systems are likely to evolve toward delivering not only predictions (for example, the risk of a deal failure) but also structured and readable justifications. Such systems may reference similar historical cases, statistical evidence, and explicitly indicate which facts and assumptions underpin their conclusions.

Research on explanation interfaces by Kim et al. and Rong et al. highlights the importance of tailoring explanations to the user. In practice, this may result in interactive dashboards for executives and managers, where market forecasts and strategic recommendations are accompanied by visualizations illustrating the underlying assumptions and key drivers. Particular attention will be paid to aligning these interfaces with managerial reasoning styles and providing appropriate abstractions, diagrams, and visual summaries.

5.3. Embedding XAI into business processes and organizational knowledge sharing

When aggregated, explanations generated by XAI systems can reveal broader organizational patterns and serve as a basis for high-level managerial insights. In this sense, XAI may accelerate bottom-up feedback: instead of lengthy reports prepared by middle management, consolidated explanations produced by AI systems can provide executives with a rapid understanding of what is happening and why.

It is conceivable that future systems will adapt to organizational culture and business specifics, learning which types of explanations decision-makers find most understandable and presenting insights in a familiar style. Complex managerial situations may require not a single recommendation, but multiple alternative scenarios, each accompanied by its own explanation. AI systems may propose several options, justify each of them, or suggest decomposing a complex problem into sub-tasks, explaining why such decomposition improves decision quality.

This leads to the concept of end-to-end explanations that cover the entire decision-making process,

from problem formulation to the final choice. At the same time, it is crucial to maintain a balance, ensuring that explanations clarify rather than distort reality. A multi-level explanation approach appears particularly promising: a concise, high-level explanation for initial understanding, complemented by a more detailed version for verification and deeper analysis.

The XAI artifacts proposed in this study (*Table 4*) are derived from an analysis of existing documentation practices, monitoring approaches, and UX methodologies, while extending them with new elements that address business needs and identified risks. As such, they represent an original contribution by the authors to the development of XAI in an organizational and business context.

Research highlights the importance of model documentation as a key mechanism for transparency and knowledge transfer across teams. Mitchell et al. [56], for example, emphasize the role of model interpretation documents such as Model Cards, while Gebru et al. [57] argue for the value of Datasheets for Datasets, which describe intended use cases, data sources, identified limitations, and mechanisms for generating explanations. The relevance of such documentation becomes particularly pronounced in large organizations, where knowledge transfer cannot rely solely on informal communication or code repositories. Well-structured documentation enables faster onboarding of new models, especially in contexts of staff turnover or solution scaling.

Equally important is the design of interaction interfaces. Whereas explanations of model outputs were previously accessible mainly through specialized analytical tools, explanations are now increasingly embedded directly into operational applications. As a result, analysts or managers can receive contextual comments alongside predictions, clarifying which factors influenced the outcome and how strongly it deviates from typical behavior. Moreover, explanation formats can be adapted to the user's role, ranging from concise business summaries to detailed technical breakdowns. Such interfaces significantly enhance AI acceptance in corporate environments, particularly when decision-making time is limited.

Table 4.

Key XAI artifacts in business contexts

Artifact	Purpose / Function	User
XAI documentation and protocols (Model Cards, Datasheets)	Formalized description of the model, data, and limitations to ensure transparency and auditability	Model developers, auditors, regulators
Explanation interfaces (Explainability UI/UX)	Interactive access to explanations within the user interface	End users, analysts, operators
XAI monitoring dashboards and panels	Visualization of factors influencing predictions to support managerial decision-making	Middle and senior management
Automated XAI validation frameworks	Automated verification of explanation quality and detection of deviations from expected behavior	Quality engineers, risk management teams, internal audit
Training and supporting materials (XAI guides, workshops)	Supporting staff in adopting and understanding XAI through training courses and guidelines	Managers, business analysts, learning and development specialists

At the strategic level of model operation, visual monitoring panels and reporting views can be used to regularly track the distributions of key features and predictions, detect data and target drift, and automatically generate alerts when thresholds are exceeded. Such monitoring supports timely identification of quality degradation and enables corrective actions (e.g., resolving data quality issues or triggering model retraining) [58].

Special attention is also given to the automated validation of explanations. In regulated domains such as banking or healthcare, manual review of every explanation is infeasible. As a result, frameworks are being developed to automatically assess explanations for compliance with internal policies, absence of discriminatory patterns, and overlooked risks. This shifts XAI from a visualization layer to an integral component of organizational quality assurance systems.

Finally, sustainable XAI adoption is impossible without educational and supporting materials accessible not only to developers, but also to a broader range of employees. These materials include guidelines,

training programs, and step-by-step instructions for interacting with XAI systems. Their purpose is to lower entry barriers and enable responsible use, particularly for professionals involved in data interpretation who may lack technical backgrounds. According to Donoso-Guzmán et al. [59], human-centered approaches to XAI evaluation that account for the goals and contexts of different user roles are especially promising. Such evaluation frameworks can also be applied to tailor explanations to corporate practices and managerial expectations, helping to anticipate which explanations will be perceived as convincing, excessive, or insufficient, and how arguments should be structured for different stakeholders.

Overall, XAI artifacts do more than merely explain individual outcomes; they facilitate the integration of AI into organizational reasoning, turning it into a transparent, accessible, and governable tool.

The outlined perspectives form a roadmap for XAI development in the coming years. The dominant trend is a deeper integration of AI and human judgment, shifting from narrowly focused explanations of indi-

vidual predictions toward human-centered intelligence embedded in collective decision-making processes. In this sense, XAI is evolving from a standalone interpretability module into a design paradigm for AI systems that are inherently conceived for collaboration with humans.

Conclusion

Explainability has become a critical requirement for AI applications in business, industry, and management, where real financial resources, safety, and responsibility toward people are at stake. It is increasingly a factor of both competitiveness and regulatory compliance: organizations that are able to explain the behavior of their algorithms are better positioned to withstand regulatory scrutiny and to gain the trust of customers and stakeholders. Even today, XAI facilitates the integration of AI into smart manufacturing and the financial sector, positively influencing the development prospects of Industry 5.0, where human-centricity, safety, and sustainability play a central role.

Explainable AI enables companies to meet these demands by providing tools for algorithm monitoring, decision rationale reporting, and control of discriminatory behavior. In the foreseeable future, XAI may become an integral part of enterprise quality manage-

ment systems. Just as ISO standards currently exist for business processes, similar standards may emerge for the explainability and ethical compliance of AI components embedded in organizational workflows.

Decision-makers equipped with algorithms that can justify their outputs gain a powerful instrument that combines the strength of data-driven models with the clarity and logical structure of traditional analysis. This enables more informed yet innovative decision-making: while AI can uncover non-obvious patterns, explainability makes these insights acceptable and actionable in practice. New forms of organizational learning based on interaction with XAI systems may further accelerate the dissemination of best practices and institutional knowledge.

At the same time, it must be acknowledged that large-scale adoption of XAI is still constrained by economic, technical, and organizational barriers. The greatest potential for XAI development lies in domains where regulatory pressure, high cost of error, and strong data and process discipline converge, such as finance, insurance, manufacturing, and industrial operations. Where organizations deliberately invest in explainability and embed it into system architectures, quality processes, and user experience design, explainable AI can become a source of sustainable competitive advantage. ■

References

1. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. (2019). In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Lecture Notes in Computer Science*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-28954-6>
2. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
3. IEEE Std 7001-2021. (2022). IEEE Standard for Transparency of Autonomous Systems. IEEE. <https://doi.org/10.1109/IEEESTD.2022.9726144>
4. ISO/IEC TR 24028:2020. (2020). Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence. Geneva: International Organization for Standardization. <https://www.iso.org/standard/77608.html>

5. Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68–77. <https://doi.org/10.1145/3359786>
6. Gamoura, S. C. (2023). Explainable AI (XAI) for AI-Acceptability: The coming age of digital management 5.0. *2023 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, 1–6. <https://doi.org/10.1109/icnsc58704.2023.10319030>
7. Khan, A., Jhanjhi, N. Z., Hamid, D. H. T. B. A. H., & Omar, H. A. H. bin H. (2024). The need for explainable AI in Industry 5.0. *Advances in Explainable AI Applications for Smart Cities*, 1–30. <https://doi.org/10.4018/978-1-6684-6361-1.ch001>
8. Chang, T.-S., & Bau, D.-Y. (2024). eXplainable artificial intelligence (XAI) in business management research: A success/failure system perspective. *Journal of Electronic Business & Digital Economics*, 4(1), 36–53. <https://doi.org/10.1108/jebde-07-2024-0019>
9. Tchuente, D., Lonlac, J., & Kamsu-Foguem, B. (2024). A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications. *Computers in Industry*, 155, 104044. <https://doi.org/10.1016/j.compind.2023.104044>
10. Martins, T., de Almeida, A. M., Cardoso, E., & Nunes, L. (2024). Explainable Artificial Intelligence (XAI): A systematic literature review on taxonomies and applications in finance. *IEEE Access*, 12, 618–629. <https://doi.org/10.1109/access.2023.3347028>
11. European Commission: Directorate-General for Research and Innovation. (2021). *Industry 5.0: Towards a sustainable, human-centric and resilient European industry*. Publications Office of the European Union. <https://data.europa.eu/doi/10.2777/308407>
12. Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, 18(8), 5031–5042. <https://doi.org/10.1109/tii.2022.3146552>
13. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/access.2018.2870052>
14. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
15. European Union. (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 (General Data Protection Regulation). *Official Journal of the European Union*, L119, 1–88. <http://data.europa.eu/eli/reg/2016/679/oj>
16. NIST AI 100-1. (2023). Artificial intelligence risk management framework (AI RMF 1.0). National Institute of Standards and Technology. <https://doi.org/10.6028/nist.ai.100-1>

17. Zavodna, L. S., Überwimmer, M., & Frankus, E. (2024). Barriers to the implementation of artificial intelligence in small and medium sized enterprises: Pilot study. *Journal of Economics and Management*, 46, 331–352. <https://doi.org/10.22367/jem.2024.46.13>
18. Weitz, K., Dang, C. T., & André, E. (2022). Do we need explainable AI in companies? Investigation of challenges, expectations, and chances from employees' perspective. *arXiv:2210.03527*. <https://doi.org/10.48550/arXiv.2210.03527>
19. Darvish, M., Kret, K. S., & Bick, M. (2024). An explorative study on the adoption of explainable artificial intelligence (XAI) in business organizations. In: van de Wetering, R., et al. *Disruptive Innovation in a Digitally Connected Healthy World*. Lecture Notes in Computer Science, 14907, 29–40. Springer, Cham. https://doi.org/10.1007/978-3-031-72234-9_3
20. Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *Npj Digital Medicine*, 6(1). <https://doi.org/10.1038/s41746-023-00751-9>
21. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080. <https://doi.org/10.1073/pnas.1900654116>
22. Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv:1708.08296*. <https://doi.org/10.48550/arXiv.1708.08296>
23. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*. <https://doi.org/10.48550/arXiv.1702.08608>
24. Yuan, H., Yang, F., Du, M., Ji, S., & Hu, X. (2021). Towards structured NLP interpretation via graph explainers. *Applied AI Letters*, 2(4), e58. <https://doi.org/10.1002/ail2.58>
25. Dumka, A., Chaudhari, V., Bisht, A. K., Rawat, R., & Pandey, A. (2024). Methods, techniques, and application of explainable artificial intelligence. In R. Gupta, A. Jain, J. Wang, & R. Pateriya (Eds.), *Reshaping Environmental Science Through Machine Learning and IoT*, 337–354. IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-2351-9.ch017>
26. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3063289>
27. Dixit, M., Kansal, I., Khullar, V., Kumar, R., & Kumar, S. (2025). Analyzing trustworthiness and explainability in artificial intelligence: A comprehensive review. *Recent Advances in Electrical & Electronic Engineering*, 18(8), article e040724231621. <https://doi.org/10.2174/0123520965308169240616144800>
28. Vasanth, S., Keerthana, S., & Saravanan, G. (2024). Demystifying AI: A robust and comprehensive approach to explainable AI. *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*, 1–5. <https://doi.org/10.1109/icec59683.2024.10837078>

29. Avdoshin, S. M., & Pesotskaya, E. Yu. (2022). Trusted artificial intelligence: Strengthening digital protection. *Business Informatics*, 16(2), 62–73. <https://doi.org/10.17323/2587-814x.2022.2.62.73>
30. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: Feasible and actionable counterfactual explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350. <https://doi.org/10.1145/3375627.3375850>
31. Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, 4, 752558. <https://doi.org/10.3389/frai.2021.752558>
32. Hjelkrem, L. O., & Lange, P. E. de. (2023). Explaining deep learning models for credit scoring with SHAP: A case study using open banking data. *Journal of Risk and Financial Management*, 16(4), 221. <https://doi.org/10.3390/jrfm16040221>
33. Li, Y., Simon, Z., & Turkington, D. (2021). Investable and interpretable machine learning for equities. *Journal of Financial Data Science*, 4(1), 54–74. <https://doi.org/10.3905/jfds.2021.1.084>
34. Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible AI. *Frontiers in Artificial Intelligence*, 5, 779799. <https://doi.org/10.3389/frai.2022.779799>
35. Watanabe, A., Kuramata, M., Majima, K., Kiyohara, H., Kondo, K., & Nakata, K. (2021). Constrained Generalized Additive 2 Model with consideration of high-order interactions (CGA2M+). *arXiv:2106.02836*. <https://doi.org/10.48550/arXiv.2106.02836>
36. IBM. (2023). *IBM Maximo Predict*. IBM Documentation. <https://www.ibm.com/docs/en/mhmpmh-and-p-u/cd?topic=overview-maximo-predict>
37. Hermansa, M., Kozielski, M., Michalak, M., Szczyrba, K., Wróbel, Ł., & Sikora, M. (2021). Sensor-based predictive maintenance with reduction of false alarms – A case study in heavy industry. *Sensors*, 22(1), 226. <https://doi.org/10.3390/s22010226>
38. Kilari, S. D. (2025). The role of explainable AI (XAI) in improving transparency and trust in supply chain demand and price forecasting models. *SSRN preprint*. <https://doi.org/10.2139/ssrn.5357669>
39. Siemens. (2023). *The rise of industrial explainable artificial intelligence (XAI) – Insights across the AI life cycle*. White Paper. <https://assets.new.siemens.com/siemens/assets/api/uuid:3b4de373-57e2-4329-b025-2825db0172aa/WhitepaperXAI.pdf>
40. Jean-Quartier, C., Bein, K., Hejny, L., Hofer, E., Holzinger, A., & Jeanquartier, F. (2023). The cost of understanding – XAI algorithms towards sustainable ML in the view of computational cost. *Computation*, 11(5), 92. <https://doi.org/10.3390/computation11050092>
41. European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L257, 1–64. <https://data.europa.eu/eli/reg/2024/1689/oj>

42. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, 57(8), 216. <https://doi.org/10.1007/s10462-024-10854-8>
43. Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, 33, 26. <https://doi.org/10.1007/s12525-023-00644-5>
44. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832. <https://doi.org/10.3390/electronics8080832>
45. Molnar C. (2025). *Interpretable machine learning. A guide for making black box models explainable*. 3rd edition. <https://christophm.github.io/interpretable-ml-book/>
46. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5), e1424. <https://doi.org/10.1002/widm.1424>
47. Rai, A. (2019). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. <https://doi.org/10.1007/s11747-019-00710-5>
48. Liao, Q. V., & Varshney, K. R. (2021) Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv:2110.10790*. <https://doi.org/10.48550/arXiv.2110.10790>
49. Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*, 11, 78994–79015. <https://doi.org/10.1109/access.2023.3294569>
50. Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4, 688969. <https://doi.org/10.3389/fdata.2021.688969>
51. d'Avila Garcez, A. S., Broda, K. B., & Gabbay, D. M. (2002). Neural-Symbolic Learning Systems. In *Perspectives in Neural Computing*. Springer. <https://doi.org/10.1007/978-1-4471-0211-3>
52. Besold, T. R., d'Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kuehnberger, K.-U., Lamb, L. C., Miikkulainen, R., & Silver, D. L. (2017) Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. *arXiv:1711.03902*. <https://doi.org/10.48550/arXiv.1711.03902>
53. Yu, D., Yang, B., Liu, D., Wang, H., & Pan, S. (2023). A survey on neural-symbolic learning systems. *Neural Networks*, 166, 105–126. <https://doi.org/10.1016/j.neunet.2023.06.028>
54. Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence*, 7. <https://doi.org/10.3389/frai.2024.1456486>
55. Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4), 2104–2122. <https://doi.org/10.1109/tpami.2023.3331846>

56. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the *Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
57. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
58. Vadapalli, S. R. (2022). Monitoring the performance of machine learning models in production. *International Journal of Computer Trends and Technology*, 70(9), 38–42. <https://doi.org/10.14445/22312803/IJCTT-V70I9P10559>
59. Donoso-Guzmán, I., Ooge, J., Parra, D., & Verbert, K. (2023). Towards a comprehensive human-centred evaluation framework for explainable AI. In: Longo, L. (eds) *Explainable Artificial Intelligence (xAI 2023)*. Communications in Computer and Information Science, 1903. Springer, Cham. https://doi.org/10.1007/978-3-031-44070-0_10

About the authors

Sergey Mikhailovich Avdoshin

Candidate of Sciences (Technology);

Professor, School of Computer Engineering, HSE Tikhonov Moscow Institute of Electronics and Mathematics, HSE University, 34 Tallinskaya St., Moscow 123458, Russia;

E-mail: savdoshin@hse.ru

ORCID: 0000-0001-8473-8077

Elena Yuryevna Pesotskaya

Candidate of Sciences (Economics);

Associate Professor, School of Software Engineering, Faculty of Computer Science, HSE University, 11 Pokrovsky Blvd., Moscow 109028, Russia;

E-mail: epesotskaya@hse.ru

ORCID: 0000-0003-2129-4645