Technological Forecasting & Social Change 229 (2026) 124710

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Opening the black box: Operational principles, tools and frameworks that advance explainable artificial intelligence (XAI) models

Mark Anthony Camilleri [a,b,c,*]

[a] University of Malta, Department of Corporate Communication, Faculty of Media and Knowledge Sciences, Imsida, MSD2080, Malta
[b] Northwestern University, Medill School of Journalism, Media, Integrated Marketing Communications, 1845m Sheridan Road, Evanston, IL, 60208, United States of America.
[c] University of Edinburgh, The Business School, 29 Buccleuch Pl, Edinburgh, EH8 9JS, United Kingdom

ARTICLE INFO

Keywords:
Ante-hoc explanations
Black box model
Counterfactual explanations
Human-in-the-loop (HITL)
Interpretable artificial intelligence
Post-hoc explanations

ABSTRACT

As artificial intelligence (AI) models are increasingly becoming permeated across various domains, there are instances where they are generating hallucinations, misinformation and erroneous outputs. Various stakeholders, particularly the regulatory ones, are encouraging the developers of machine learning (ML) systems to clarify or justify their models' decisions, actions or predictions in a way that is understandable to their users. In this light, this article raises awareness on Explainable Artificial Intelligence (XAI) principles that are intended to increase transparency, accountability and fairness about the modus operandi of machine learning algorithms. A systematic review of the extant literature identifies key tools, frameworks and best practices that enhance the interpretability of AI models, including open-source techniques like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), among others. The synthesis of the findings also shed light on XAI challenges and limitations of black-box models. This contribution advances a conceptual framework for the responsible implementation of XAI and offers practical guidelines that promote the interpretability of AI systems, whilst addressing their opacity, as well as their biased outcomes. It puts forward theoretical and managerial implications as well as future research avenues.

1. Introduction

Explainable artificial intelligence (XAI) is a popular notion that is gaining traction among various stakeholders, as practitioners who are researching, developing and maintaining machine learning systems are encouraged to adopt a responsible approach to AI models (Dwivedi et al., 2023; Yeo et al., 2025). Essentially, XAI is related to the promotion of transparent, accurate and fair models, that build trust and confidence in machine learning capabilities (Lundberg et al., 2020; Singu et al., 2026; Wang et al., 2024). This concept is emerging as more users of AI-driven technologies are questioning how their algorithm works and how they generate their results. Various stakeholders are concerned about potential AI errors (Bao et al., 2025). In the past, there were instances where AI training models were unfair, biased and predisposed, either positively or negatively towards certain individuals, based on their race, gender, age or on the location where they reside (Deepa et al., 2024). This happens because there could be a distribution shift between training data and production data. Errors may arise because there are instances when real-world data may not match the model's expectations.

For example, a model's performance can be compromised if dynamic production data is collected from real-world environments. The data gathered can be noisy, inconsistent, missing or outdated, because conditions continuously change. Production data may fail to capture the latest patterns, as user behaviors, contexts, and/or external environments evolve. As a result, the models based on such data reflect shifting distributions (concept drift), leading to erratic or biased outputs, that are usually considered as unreliable or unfair. On the other hand, static training data is fixed at the moment it is collected. In this case, the data can be outdated, but is not erratic, because it does not change over time. Any incompleteness or bias in the data is embedded at the point of collection rather than emerging from ongoing, dynamic processes (Herm et al., 2023; Janssen et al., 2022). Hence, it is imperative to continuously monitor model drift, as there may be changes in data or changes in relationships between input and output variables (Hinder et al., 2023; Rahmani et al., 2023). Models should be updated as frequently as possible by using online learning or by retraining pipelines in order to

* Corresponding author at: University of Malta, Department of Corporate Communication, Faculty of Media and Knowledge Sciences, Imsida, MSD2080, Malta.
E-mail address: mark.a.camilleri@um.edu.mt.

https://doi.org/10.1016/j.techfore.2026.124710
Received 18 June 2025; Received in revised form 3 March 2026; Accepted 23 April 2026
Available online 29 April 2026

maintain robust AI systems (Salem et al., 2024).

AI's calculation processes may not be always clear, even to the scientists or engineers who developed the algorithm, let alone to typical, non-expert users. At times, systems administrators consider machine learning technologies as "black boxes" as they cannot explain or interpret the algorithms' methodologies that they employed to obtain responses, results or outputs (Li, 2022). The black box concept suggests that practitioners are not always in a position to fully understand AI's decision-making processes (Adadi and Berrada, 2018). Yet, they ought to be accountable for the machine learning (ML) algorithms, for deep learning (Adak et al., 2022; Askr et al., 2023; Ciravegna et al., 2023; Yang et al., 2022) and for the neural networks (Ciravegna et al., 2023; Confalonieri et al., 2021; Garcez and Lamb, 2023; Wang et al., 2021; Yang et al., 2022) they create.

Practitioners are increasingly referring to Open-Source Tools and Frameworks, including to SHapley Additive exPlanations (SHAP) (Barnard et al., 2022; Li, 2022; Yang et al., 2024; Zhang et al., 2022) and to Local Interpretable Model-agnostic Explanations (LIME) (Adak et al., 2022; Shams Amiri et al., 2021) to better understand the output of machine learning models. While SHAP uses game theory to explain the prediction of a data sample by calculating the contribution of each feature to the prediction of the algorithm (Baptista et al., 2022; Fan et al., 2024), LIME involves a technique that approximates any black box machine learning model with a local, interpretable model, to shed light on each individual prediction (Shams Amiri et al., 2021). Practitioners can avail themselves of other resources like EL15 (i.e. a Python library), that can help with the visualization and debugging of machine learning classifiers. EL15 explains their predictions, implements algorithms and offers unified application program interfaces (APIs) for various model types including black box models.

In addition, practitioners may refer to Fiddler AI's platform that is designed to enhance the transparency, reliability and accountability of machine learning (ML) and large language model (LLM) applications (Cambria et al., 2023). Furthermore, they could access research papers published through leading AI conferences like the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Machine Learning (ICML) as well as the Association for the Advancement of Artificial Intelligence (AAAI), among others, in order to acquaint themselves with cutting-edge XAI research.

This research presents the findings from a systematic review of the extant literature focused on XAI. Specifically, its research objectives are threefold: Firstly, it scrutinizes high impact contributions in their entirety. It evaluates their research questions, methodologies, key findings and implications, to identify XAI practices, tools and frameworks that are intended to improve the machine learning systems' transparency, reliability and accountability. Therefore, it investigates the role of open-source tools (e.g., SHAP, LIME, ELI5) and commercial platforms (e.g., Fiddler AI) in improving the interpretability of AI model outputs in varied application domains.

Secondly, this paper explores the ethical issues arising from black-box AI models, particularly their lack of clarity in high-stakes applications. It discusses operational principles and frameworks and discusses about the challenges related to their implementation. Hence, it provides a comparison matrix that clearly outlines XAI tools and specifies their metrics. It summarizes their strengths, weaknesses/limitations and domain fit.

Thirdly, this research puts forward a conceptual framework and substantive guidelines that are anchored in influential key theoretical underpinnings related to "Responsible Research and Innovation" (EU, 2014; Owen et al., 2021; Stahl and Wright, 2018) and to "Technical Systems Theory" (Appelbaum, 1997; Trist, 1981), that are ultimately intended to foster responsible governance through XAI processes under conditions of technological uncertainty. This framing is especially powerful if XAI is connected to anticipatory governance, early warning signals or to reflexive monitoring in evolving AI regimes. For instance, the Socio-Technical Systems Theory (STS) views AI systems as co-evolving configurations of technology, institutions, actors and norms. From this perspective: XAI matters because opacity disrupts alignment between the technical subsystem and the social subsystem. Similarly, the Responsible Research and Innovation (RRI) recommendations emphasize anticipation, reflexivity inclusion and responsiveness as preconditions for responsible AI innovations. Hence, the RRI policy paradigm is very pertinent for the development and growth of XAI systems, as it enables ex ante impact assessments, scenario analysis, regulatory sandboxes, and dynamic risk classification.

In this light, this research clarifies how, why and when practitioners integrate tools and methodologies that promote the explainability of AI systems, from initial problem definition and data collection to model training, evaluation, deployment, as well as ongoing monitoring and maintenance. Practitioners are expected to monitor data drift, and to ensure fairness and accountability across their AI products' lifecycles. The proposed conceptual framework's underlying goal is to raise awareness on the importance of transparent, trustworthy and fair XAI systems, for the benefit of users who are availing themselves of these technologies, as well as for practitioners who are involved in their research, development and maintenance.

This article contributes to the ongoing discourse on responsible AI governance and addresses gaps in prior XAI research by integrating technical, ethical and data-centric perspectives. Comprehensive synthesis of recent XAI research, evaluates current tools and practices, and proposes a structured set of insights for practitioners. It provides both theoretical and practical lenses to increase the stakeholders' knowledge about how XAI can be operationalized in various machine learning applications in different contexts. It adds value to the extant literature on XAI, as it provides a clear analytical lens focused on ante-hoc vs. post-hoc of XAI approaches rather than merely describing its taxonomies. It discusses perennial challenges in AI explainability research like model drift as well as on shifts in production data (data drift), that could undermine the reliability and fairness of AI systems. Such issues are often overlooked in the extant literature (Hinder et al., 2023; Rahmani et al., 2023). In addition, it assesses real-world tools (e.g., SHAP, LIME, ELI5, Fiddler AI), and proposes a lifecycle-based, multi-stakeholder framework that links explainability practices to responsible AI governance, across dynamic application contexts. In sum, this research proposes a multi-stakeholder view of XAI that emphasizes not only the needs of developers but also considers the expectations of system administrators and policymakers, as well as of non-expert users.

2. Methodology

This research uses a grounded theory methodology involving systematic data gathering processes and substantive analyses of high impact academic publications. The study identified relevant sources focused on XAI. A systematic review protocol was used to search, screen, extract and synthesize the findings from Scopus-indexed articles.

An open-ended temporally unconstrained query focused on the terms "explainable artificial intelligence" or "explainable AI" or "XAI" that was limited to the title, abstract and keywords fields, specifically, "TITLE-ABS-KEY ("Explainable artificial intelligence" OR 'Explainable AI' OR 'XAI')" reported that there were 25,240 documents that featured one or more of these keywords.

Hence, the search was narrowed down to identify only articles published in English speaking journals from 2020 onwards. The query comprised the following information: "TITLE-ABS-KEY ("Explainable artificial intelligence" OR "Explainable AI" OR "XAI") AND PUBYEAR > 2019 AND PUBYEAR < 2027 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (SRCTYPE, 'j'))" yielded 10,915 documents.

Subsequently, these documents were reduced to 940 documents, when this query focused on publications related to social sciences. Specifically, the search string involved the following entry through Scopus: "TITLE-ABS-KEY ("Explainable artificial intelligence" OR

"Explainable AI" OR "XAI") AND PUBYEAR > 2019 AND PUBYEAR < 2027 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (SRCTYPE, "j")) AND (LIMIT-TO (SUBJAREA, 'SOCI'))".

This research features the search strings that were used in this systematic review exercise to enhance the credibility, dependability and confirmability of this research, in line with Guba and Lincoln's (1985) criteria for trustworthiness. It clearly specifies the inclusion and exclusion criteria that were adopted for this systematic review. It sheds light on the screening processes that led to the analysis of sixty ($n = 60$) of the most cited articles, as shown in Fig. 1. It sought to ensure that the final corpus of the extracted literature is both methodologically sound and theoretically valuable, thereby strengthening the validity of the grounded theory analysis and the robustness of the resulting theoretical contributions.

Table 1 features a complete list of sixty (60) of the most cited articles (sorted in alphabetical order, according to the authors' surnames). It appraises the contributing authors, clarifies their research objectives and describes the methodologies they used for their studies.

The analysis reported that most researchers adopted experimental analytical approaches to investigate XAI. Other methods included literature reviews, conceptual or discursive articles, case studies, mixed methods (quantitative and qualitative techniques), and qualitative studies. An inductive approach was employed to synthesize the most salient findings from the extracted articles. All papers were scrutinized in their entirety to identify the themes of study and to address the underlying research questions of this contribution.

3. Results

3.1. XAI operational principles, tools and frameworks

XAI has emerged as a critical area of study of AI research. This may be due to the increased stakeholders who are exerting their pressure on practitioners to become as accountable as possible, during the development and maintenance of their AI models (Langer et al., 2021; Minh et al., 2022). XAI concepts span from foundational notions in artificial intelligence and machine learning, to specialized constructs such as interpretability, transparency, and human-centered design (Gunning et al., 2019; Linardatos et al., 2020). Additionally, XAI research is informed by insights from human-computer interaction and decision science, which ultimately emphasize user engagement and trust (Chen et al., 2023; Shin, 2021). Appendix A presents concise definitions of the core terminology underpinning XAI. It offers a conceptual grounding for scholars, practitioners and regulatory stakeholders seeking to enhance their knowledge and understanding of this evolving field. The findings from this review exercise indicated that they are genuinely concerned about the complexity and opacity of modern AI systems, as they are aware that AI technologies are being integrated into critical decision-making environments, ranging from healthcare or medical systems to finance, legal and/or public administration.

This systematic review confirms that stakeholders are expecting

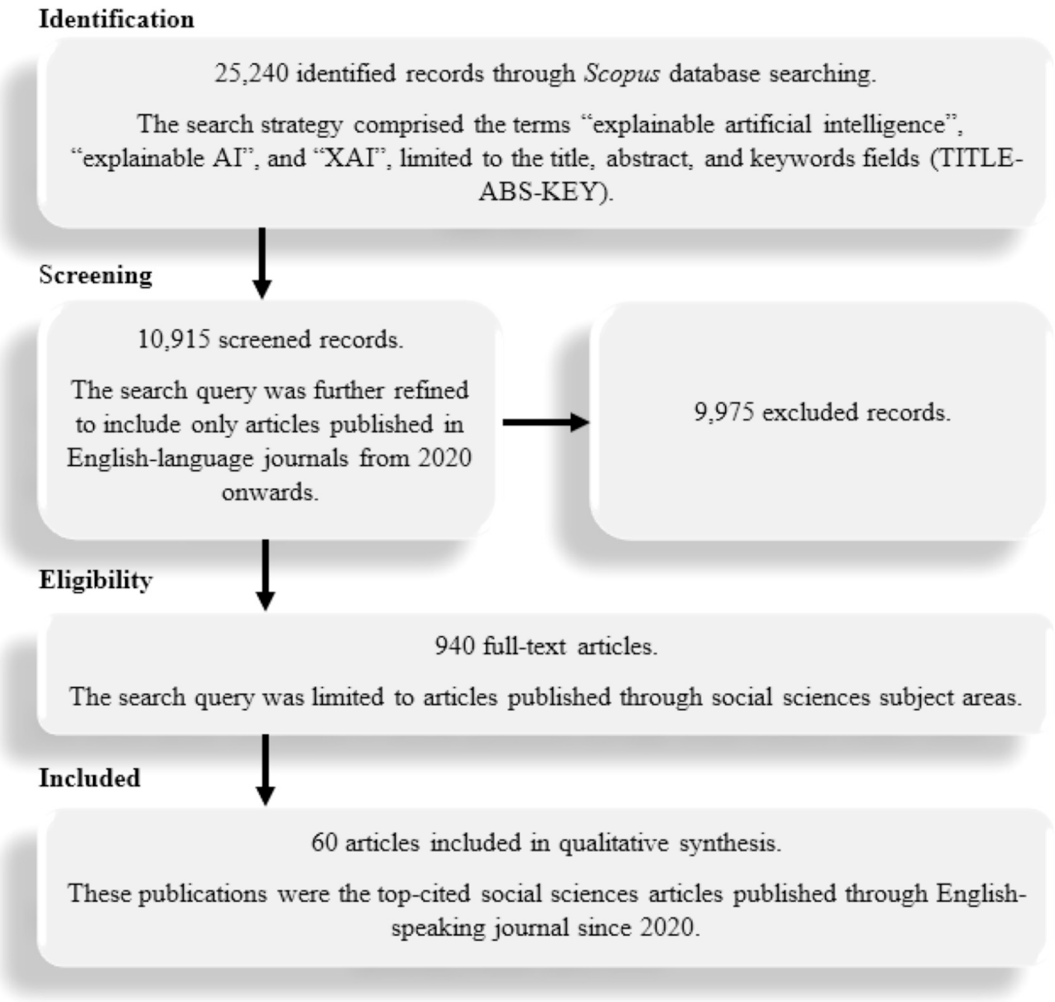

Fig. 1. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol that was adopted for this study (Adapted from Moher et al., 2009).

Table 1
A list of the most cited publications on XAI (sorted in alphabetical order).

Authors	Year	Research objective	Methodology
Adak et al., 2022	2022	This research relies on explainable artificial intelligence (XAI) methods that can evaluate customer sentiments in their online reviews of food delivery services.	Review
Andrews et al., 2023	2023	This research advances a review focused on knowledge structures of shared mental models of team members consisting of human and AI agents.	Review
Askr et al., 2023	2023	This research presents an overview of how deep learning technologies and explainable AI could support researchers in drug discovery.	Review
Baptista et al., 2022	2022	This research examines XAI's data-driven attributes and their prognostic characteristics.	Experimental
Barnard et al., 2022	2022	This research investigates how experiments conducted on NSL-KDD dataset can detect new attacks encountered during testing.	Experimental
Bauer et al., 2023	2023	This research explores the effects of feature-based AI interpretations on their users' processing of its outputs.	Experimental
Buçinca et al., 2021	2021	This research analyzes the users' reliance on explainable AI recommender systems.	Experimental
Cambria et al., 2023	2023	This research examines AI's readiness to provide natural language explanations to their users.	Review
Chen et al., 2023	2023	This research explores feature-based as well as example-based AI explanations.	Mixed methods
Ciravegna et al., 2023	2023	This research investigates how deep learning networks can learn during training to provide explanations in unsupervised contexts.	Experimental
Conati et al., 2021	2021	This research examines explainable AI-driven hints and their provision of personalized feedback (based on personality traits and cognitive abilities) in intelligent tutoring systems (ITS).	Experimental
Confalonieri et al., 2021	2021	This research utilizes an algorithm that extracts surrogate decision trees from black-box models to provide explanations and improve the understandability of AI.	Experimental
Dazeley et al., 2021	2021	This research sheds light on XAI's comprehensive explanations that their users expect out of them.	Review (conceptual)
De Bruijn et al., 2022	2022	This research uses three illustrative cases to investigate societal acceptance of explainable, data-driven decisions of AI.	Case studies
Di Martino and Delmastro, 2023	2023	This research reviews articles focused on explainable AI methods, to raise awareness of their relevance in healthcare domains.	Review
Dikmen and Burns, 2022	2022	This research explores users' trust in an AI assistant's recommendations in the context of financial investing.	Review
Garcez and Lamb, 2023	2023	This research describes the constituent parts of neurosymbolic AI systems. It focuses on studies that integrate ML with logical reasoning and knowledge.	Review
Ghaffarian et al., 2023	2023	This research delves into the burgeoning domain of explainable AI disaster risk management. It identifies various hazards, disaster types and risk components, to provide insights for decision-making.	Review
Graziani et al., 2023	2023	This research explains key terms related to the interpretability of AI systems for technical developers, academia and policymakers.	Case study / Round-table public meeting
Guidotti, 2021	2021	This research proposes an approach that clarifies the extent to which AI explanations are correct.	Experimental
Guleria and Sood, 2023	2023	This research utilizes machine learning-based white and black box models to analyze students' academic and employability attributes that are important for job placements and skilling.	Experimental
Herm et al., 2023	2023	This research evaluates the tradeoff between machine learning algorithms performance and explainability. It aims at enhancing knowledge about the decision logic of black box models.	Experimental
Holzinger, 2021	2021	This research explores explainable AI and multi-modal causability in the medical domain.	Review (discursive)
Iban, 2022	2022	This research examines how explainable AI methods can be utilized for real estate appraisal purposes.	Experimental
Jang et al., 2022	2022	This research predicts the students' performance levels in their learning environment by using ML and XAI.	Mixed methods
Janssen et al., 2022	2022	This study investigates human decision making with and without the use of ML algorithms.	Experimental
Jiménez-Luna et al., 2021	2021	This research explored the transparency of rational molecular design by training machine learning technologies to predict plasma protein binding, hERG channel inhibition, passive permeability, and cytochrome P450 inhibition	Experimental
Johnson et al., 2022	2022	This research examines human-AI interactions in order to develop informed AI, to help improve service quality and customer satisfaction.	Experimental

(continued on next page)

Table 1 (*continued*)

Authors	Year	Research objective	Methodology
Joung and Kim, 2023	2023	This research proposes an interpretable machine learning-based approach to segment customers of new products, based on their perceptions of the products' features and attributes, from online product reviews.	Sentiment analysis
Kenny et al., 2021	2021	This research uses post-hoc explanations to investigate black-box classifiers. The researchers compare a deep learning black box model with a more transparent, proxy model.	Experimental
Khosravi et al., 2022	2022	This research presents an explainable AI model. The researchers evaluate its key aspects to better design and develop educational tools.	Case studies
Laato et al.	2021	This research explores how machine learning models behave, and the basis of their outputs.	Review (discursive)
Langer et al., 2021	2021	This research raises awareness on the key stakeholders that are calling for explainable artificial intelligence systems.	Review (discursive)
Li, 2022	2022	This research involves simulation experiments that compare SHapley Additive exPlanations- (SHAP-) explained eXtreme Gradient Boosting (XGBoost) to Spatial Lag Model (SLM) and Multi-scale Geographically Weighted Regression (MGWR), at the parameter level.	Experimental
Meske et al., 2022	2022	This research describes exemplary risks of black-box AI and discusses about the consequent need for explainability of AI in information systems research.	Discursive
Minh et al., 2022	2022	This research assesses various explainable AI methods, and groups them into pre and post modeling interpretable systems.	Review (discursive)
Mosqueira-Rey et al., 2023	2023	This research focuses on human-in-the-loop (HITL) machine learning, as it sheds light on the interactions between humans and machine learning algorithms.	Review (discursive)
Naiseh et al., 2023	2023	This research examines scenarios, explainable AI interfaces and task ratings, to enhance trust calibration and improve decision-making of clinical decision support systems.	Qualitative (semi-structured interviews)
Olson et al., 2021	2021	This research focuses on the generation of counterfactual explanations of deep reinforcement learning agents that operate in visual input contexts.	Experimental
Oyebode et al., 2023	2023	This research explores the current trends in machine-learning based adaptive systems for different health and well-being applications.	Review
Panigutti et al., 2021	2021	This research discusses about a tool that can audit a fictional blackbox model. The authors suggest that it can identify biases in patient data within clinical systems.	Experimental
Rahman et al., 2021a, 2021b	2021	This research raises awareness about secure, private and explainable "internet of health things" applications that can support healthcare systems.	Experimental
Riveiro & Thill	2021	This research explores end user expectations of XAI systems.	Experimental
Sanneman and Shah, 2022	2022	This research sheds light on relevant human factors related to XAI research. The authors put forward recommendations and performance metrics for XAI systems.	Review
Schoonderwoerd et al., 2021	2021	This research discusses about the human-centered design approaches for ML-generated medical diagnoses.	Case study
Setzu et al., 2021	2021	This research focuses on rule-based local explanations. The authors raise awareness on logical explanability, local explainability and explanation composition.	Experimental
Shams Amiri et al., 2021	2021	This research advances a transportation energy model (TEM) that uses XAI methodologies to forecast energy consumption.	Case study
Shin, 2021	2021	This research investigates the users' experiences with AI algorithms. It sheds light on the impacts of explainable AI explainability on the users' trust in AI systems.	Quantitative
Susnjak et al., 2022	2022	This research investigates learning analytics dashboards (LADs) in terms of their descriptive, predictive and prescriptive analytics capabilities at a higher educational institution. The authors identify their strengths. Weaknesses and challenges in developing student-facing LADs.	Review
Tiddi and Schlobach, 2022	2022	This research raises awareness on knowledge graphs as tools for explainable AI. The authors identify the strengths as well as the limitations of such systems.	Review (discursive)
van der Waa et al., 2021	2021	This research compares rule-based and example-based AI explanations.	Experimental
Vasconcelos et al., 2023	2023	This research explores the users' overreliance on AI with and without explanations for its predictions.	Quantitative
Wang et al., 2021	2021	This research develops a generalizable chatbot architecture (CASS) based on advanced neural network algorithms in an online health service setting.	Experimental
Wysocki et al., 2023	2023	This research explores the healthcare professionals' attitudes towards explainable machine learning models for clinical decision support.	Quantitative

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Research objective	Methodology
Yang et al., 2024	2024	This research utilizes interpretable ML technologies to investigate synergistic effects of creating green spaces in a Chinese city.	Quantitative
Yang et al., 2021	2021	This research discusses the pros and cons of using AI to improve the productivity of their tasks and activities, including in the realms of education.	Review (discursive)
Yang et al., 2022	2022	This research develops a framework that can predict traffic accident severity risks with good accuracy in order to implement countermeasures and increase road safety.	Experimental
Yigitcanlar et al., 2021	2021	This research appraises the use of AI in smart cities. It evaluates practices, developments, trends and applications that address operational efficiencies, and sustainability aspects in urban contexts.	Review (discursive)
Zhang et al., 2022	2022	This research refers to explainable AI (XAI) that clarify why power system models make certain decisions. The authors suggest that SHapley Additive exPlanations (SHAPs), also known as Deep-SHAP method can provide an interpretable deep reinforcement learning (DRL) model that increase transparency and trust in the AI processes.	Review (discursive)

practitioners to develop explainable AI systems, that are not only accurate, but also interpretable, transparent and trustworthy (Khosravi et al., 2022). The findings suggest that XAI seeks to bridge the gap between technical performance and human understanding, by providing meaningful explanations for outputs generated by machine learning models (Mosqueira-Rey et al., 2023), especially for those that function as "black boxes". Several commentators indicated that XAI aims to foster user trust, supports accountability and ensures ethical and regulatory compliance (Khosravi et al., 2022; Schmidt et al., 2020).

The findings from this study confirm that the growing use of ML in sensitive areas like healthcare, finance, education and employment has sparked the stakeholders' concerns over the opacity aspects of black boxes and on the possible liabilities of practitioners who research, develop and maintain AI-driven solutions (Setzu et al., 2021; Zhang et al., 2022). Generally, XAI practices can be divided into two broad categories: (i) inherently interpretable models, such as decision trees and linear regressions, and (ii) post-hoc interpretability methods for more complex black-box models like deep neural networks (Confalonieri et al., 2021; Kenny et al., 2021). The latter one generates both local and global explanations through feature attribution, perturbation analysis and visualizations.

For the time being, several complex ML models operate as black boxes, as they hinder the ability of their users and regulators to understand, contest or improve their outputs. XAI addresses these contentious issues by providing tools, methodologies and frameworks that are intended to enhance the interpretability of AI systems through substantive compliance mechanisms, ethical standards and normative guidelines.

Indeed, this research indicates that inherently interpretable models, counterfactual reasoning, ongoing fairness audits, human-in-the-loop (HITL) approaches as well as post-hoc explanations may contribute to improving the transparency and trustworthiness of ML algorithms (Mosqueira-Rey et al., 2023; Panigutti et al., 2021). Whilst counterfactual explanations enable practitioners to explore "what-if" scenarios and offer them actionable insights that can improve their model's comprehensibility for decision subjects; regular fairness audits could analyze model outcomes across demographic groups and may possibly identify potential biases (Holzinger, 2021). In addition, human-in-the-loop (HITL) approaches and post-hoc explanations (as well as retrospective interpretability techniques) can enhance contextual accuracy and accountability.

Practitioners may avail themselves of a range of tools and libraries to implement XAI techniques including open-source options like SHAP, LIME, ELI5 and Alibi, among others, that offer model-agnostic interpretability (Baptista et al., 2022; Li, 2022; Zhang et al., 2022). Moreover, they may use IBM's AIX360 and Microsoft's InterpretML to support them in the explainability of their datasets and machine learning models throughout their AI application lifecycle. Both resources include a diverse set of algorithms, code, guides, tutorials and demos that can help users better understand and explain AI models. Furthermore, they may utilize Google's What-If Tool (WIT), an interactive visual interface designed to help data scientists, machine learning practitioners and AI ethicists explore, analyze and explain ML models. Such tools enable non-experts, researchers and practitioners to assess model fairness, evaluate performance, deploy responsible systems and make alternative predictions (Camilleri, 2023; Herm et al., 2023; Jang et al., 2022).

Currently, there are a number of XAI frameworks and evaluation standards that can institutionalize transparency. For example, these include initiatives like the United States' Government's Defense Advanced Research Projects Agency (DARPA) XAI Program whose AI systems' decision-making processes can be understood and trusted by humans (Malach et al., 2025). DARPA funded a variety of interdisciplinary teams included academia, industry and national labs that explored human-AI interactions (interfaces and feedback loops to improve user trust and usability), interpretable ML models, post-hoc visual and symbolic explainable methods for black-box models like deep neural networks, cognitive psychology integration that design explanations that align with how humans reason and make sense of information. Similarly, Microsoft's Prediction-Decision-Recommendation (PDR) framework offers an operational and predictive model for building trustworthy AI recommendation systems that are aligned with human values. PDR was introduced as part of Microsoft's efforts in responsible AI, particularly in enterprise and applied settings.

Both DARPA's XAI Program and Microsoft's PDR (Prediction-Decision-Recommendation) framework can incorporate both quantitative and qualitative assessments in their XAI evaluation. For example, DARPA-funded XAI projects quantitative assessment involves an examination of the models' (i) Fidelity of their logic, (ii) Completeness, and (iii) Simplicity. They use performance metrics to evaluate task accuracy, latency or robustness under explanation constraints. The qualitative assessment emphasizes human-centered evaluation as it investigates perceptions about task effectiveness as well as user trustworthiness, user expectations, and user satisfaction levels with AI models.

Furthermore, standards such as IEEE P7003 Standard for Algorithmic Bias Consideration that is part of the Institute of Electrical and Electronics Engineers (IEEE) P7000 series of standards for Ethically Aligned Design in autonomous and intelligent systems (AIS) is aimed at providing technical guidance for identifying, documenting and mitigating algorithmic bias in AI systems throughout their design, development and deployment. Other tools like Fairlearn and Testing with Concept Activation Vectors (TCAV) (a post-hoc explainability method) help assess model behavior against abstract social concepts. They intend to assist developers and data scientists in assessing and improving the fairness of ML technologies including ensemble methods and deep neural networks (Confalonieri et al., 2021; Kenny et al., 2021).

3.2. XAI challenges and methodological limitations

While deep learning infrastructures like black-box AI models often exhibit remarkable predictive performance, they suffer from a lack of interpretability, as it is difficult to understand the internal logic or rationale behind their decision-making processes and predictions (Adadi and Berrada, 2018; Shajalal et al., 2024). The lack of transparency and trustworthiness of black box models undermines the stakeholders' efforts to audit or assign accountability for model-driven actions (Panigutti et al., 2021). It may prove hard to ensure that AI developers and systems administrators are held answerable for model-driven actions, when and if errors or harm occur, especially in certain industry sectors like healthcare diagnostics, financial aspects like credit-scoring, and/or welfare allocations, among others. In these contexts, their ML technologies' decisions may have profound and potentially irreversible effects on the individuals' lives. Without insight into how decisions are made, affected parties have limited avenues for recourse to action or to equitable remedies, thereby undermining procedural fairness, that could lead to the erosion of public trust in algorithmic systems.

ML systems are typically trained on datasets that may embed historical or structural biases, thereby posing risks of perpetuating inequitable outcomes in automated decision-making (Panigutti et al., 2021). This may result in a situation where a decision-making process or algorithm disproportionately and negatively affects vulnerable or underrepresented groups in society, even if there is no explicit intent to discriminate against them, particularly if their data may be undersampled or misrepresented in training sets.

AI models' predictive accuracy and fairness may degrade over time due to effects of data drift on the performance of machine learning models (Hinder et al., 2023; Rahmani et al., 2023). Shifts in the underlying data distribution or changes in real-world contexts (e.g. political, economic, social, technological and/or ethical issues) can cause the models to produce less reliable or biased outcomes, thereby necessitating continuous monitoring, periodic retraining and fairness audits to ensure sustained performance and regulatory compliance (Yang et al., 2021). Such changes may either occur gradually (concept drift) or abruptly (covariate shift). Consequentially, models trained on historical data may no longer generalize well. Their ML systems may yield suboptimal outcomes that can impact on the livelihoods of individuals and specific groups in society. For instance, a financial institution that relies on a credit-scoring model that was trained before major economic fluctuations (e.g. inflation, recession and/or rises in taxes, duties and tariffs) could penalize individuals from economically disrupted regions without accounting for recent changes in income dynamics (Dikmen and Burns, 2022). Alternatively, low-income or minority borrowers including single mothers, immigrants or disabled persons (among other vulnerable groups in society) could be denied fair access to bank credit, as AI systems may fail to reflect new socioeconomic changes in the labor market. As a result, AI systems risk perpetuating or exacerbating existing inequalities without adequate and sufficient mechanisms that ensure that AI systems are fair and up to date with the latest developments.

It is imperative that AI practitioners conduct fairness auditing on a regular basis (Panigutti et al., 2021). They need to evaluate and appraise algorithmic outputs across various demographic groups to identify and correct any disparate impacts. Such audits must go beyond one-time assessments and need to become an integral part of the AI lifecycle, in order to ensure that models evolve in ways that uphold ethical standards and regulatory requirements (Buçinca et al., 2021). When combined, explainability, monitoring and fairness auditing can establish a trustworthy AI that is clearly aligned with societal expectations of justice, equity and accountability.

Indeed, XAI techniques can help address ethical and performance-related concerns by providing transparency into model behavior, as stakeholders including regulatory bodies, AI developers, auditors and affected individuals have a legitimate right to understand how specific outcomes are generated. Practitioners who maintain AI systems ought to regularly monitor the models to identify early warning signs of degradation. They are expected to recalibrate them before harmful consequences arise (Sun et al., 2023).

AI practitioners are encouraged to advance interpretable and efficient models that are responsive to the diverse needs of end users including data scientists, domain experts and end-users. Their human-centred evaluation of XAI methods may usually focus on the development of comprehensible explanations. Hence, they refer to common metrics including: (i) sparsity (meaning that explanations highlight only the most important factors) (Guleria and Sood, 2023); (ii) explanation complexity (referring to how simple or complicated an explanation is) (Johnson et al., 2022); (iii) simulatability (which is the extent to which practitioners can anticipate the model's decision after seeing the explanation) (Vasconcelos et al., 2023); and (iv) coverage, that indicates how many cases an explanation applies to) (De Bruijn et al., 2022). In addition, user-centered outcomes such as trust in the system, improved task performance and the time required to understand the explanation are also considered by AI administrators (Buçinca et al., 2021; Dikmen and Burns, 2022; Naiseh et al., 2023; Wysocki et al., 2023).

More importantly, their systems ought to be legally and ethically justifiable as well as socially defensible. They are required to comply with relevant regulatory frameworks governing the deployment of their models and to meet the transparency and auditability standards set by specific jurisdictions, such as the European Union's General Data Protection Regulation (GDPR) and its AI Act (2024), among others. Table 2 features a comparison matrix that provides a non-exhaustive list of XAI tools. It outlines their strengths, weaknesses / limitations and identifies potential domains in which these tools can be applied.

3.3. A conceptual framework for XAI

Explainable artificial intelligence (XAI) has become a very important area of inquiry for the promotion of responsible AI governance. Regulators, organizations and end-users are increasingly demanding that ML systems are transparent, accountable and fair. Beyond technical performance, these technologies are now expected to protect users' privacy, safety and security, while remaining inclusive and accessible for the benefit of diverse socio-demographic groups in society, regardless of their age, gender, ability or ethnicity. As a result, XAI is no longer a peripheral consideration; rather, it has become a normative requirement as it advances ethical, trustworthy, and socially legitimate AI systems.

Accordingly, the objectives of XAI, whether explainable ML designs are driven by regulatory compliance, operational transparency policies or for trust building purposes, ought to be embedded across the entire AI lifecycle. The explainability of AI plays a critical role during the research and development phase, from data collection and preprocessing to model training, deployment, monitoring and maintenance. Notwithstanding, AI systems are better positioned to achieve accountability, reliability, and ethical alignment, when explainability is treated as an integral component of process innovation rather than a retrospective add-on,

However, there are instances during model development, where practitioners may have to balance trade-offs between the predictive performance of AI systems and their interpretability. Hence, evaluation criteria need to extend beyond accuracy and efficiency. They should consider the extent to which models generate explanations that are meaningful, accessible and appropriate for different user groups. Therefore, data-related practices are particularly influential at this stage. Transparent data provenance, systematic bias auditing as well as input features that are presented in a way that are easily understandable manner to humans (i.e. human-readable feature engineering) can substantially enhance model interpretability and user trust. In this respect, inherently interpretable models, such as decision trees and generalized additive models (GAMs) offer direct insights into decision logic, in contrast to complex black-box models that rely on post-hoc explanation techniques.

Table 2
A comparison matrix of XAI tools that specifies their key metrics, strengths, weaknesses/limitations and domain fit.

XAI tool	Type	Core metrics	Supporting / human-centered metrics	Strengths	Weaknesses / limitations	Possible domains
Inherently interpretable models (decision trees, linear/logistic regression, rule-based models)	Model class	Sparsity Explanation length / complexity Rule length Simulatability Time-to-understanding	Coverage (model-wide) User trust score	- Transparent, easy to explain - Supports regulatory compliance - High interpretability without post-hoc tools	- Limited predictive power for complex patterns - May oversimplify high-dimensional data	- Finance (credit scoring) - Public administration - Healthcare triage - Education and HR screening
Post-hoc interpretability (general category)	Methodological class	Explanation length / complexity Sparsity Coverage Visualization clarity	User trust score Time-to-understanding	- Allows explanation of black-box models - Generates local and global explanations - Broad domain applicability	- Risk of misleading explanations - Does not make the model itself interpretable - May be computationally intensive	- Deep learning applications - High-stakes decisions needing model transparency
Counterfactual explanations	Method	Sparsity Explanation length Time-to-understanding User trust score	Coverage Task performance improvement	- Intuitive "what-if" reasoning - Actionable for decision subjects - Enhances user agency and contestability	- May propose unrealistic or infeasible scenarios - Sensitive to feature correlations	- Finance (loan decisions) - Hiring & admissions - Healthcare prognosis
Fairness audits (ongoing)	Governance mechanism	Coverage Visualization clarity	User trust score Task performance improvement	- Detects structural biases - Essential for compliance (e.g., EU AI Act, GDPR) - Supports trust and equity	- Requires access to sensitive demographic data - Needs continuous monitoring - May uncover issues that require costly remediation	- Public sector decision systems - Finance (credit scoring) - Policing algorithms - Welfare allocation
Human-in-the-loop (HITL)	Operational approach	User trust score Task performance improvement Time-to-understanding	Visualization clarity Explanation length	- Enhances accountability - Reduces automation bias - Supports hybrid decision-making	- Slows automation - Human reviewers require training - May introduce human bias	- Healthcare diagnosis - Legal assessments - Safety-critical systems
SHapley Additive exPlanations (SHAP)	Post-hoc; model-agnostic	Sparsity Explanation length / complexity Coverage Visualization clarity	Simulatability Time-to-understanding	- Theoretically grounded (game theory) - Local and global explanations - Widely adopted, rich visualization tools	- High computational cost for large models - Can overwhelm non-experts with detail	- Tabular/ structured data - Finance, insurance, healthcare
Local interpretable model-agnostic explanations (LIME)	Post-hoc; model-agnostic	Sparsity Explanation length Coverage	Simulatability Time-to-understanding	- Simple, intuitive local explanations - Lightweight and fast - Works across model types	- Instability of explanations - Locality sampling may be misleading	- Real-time decisions - Early-stage diagnostics of ML models
ELI5	Model-agnostic toolkit	Sparsity Explanation length Visualization clarity	Simulatability	- Easy-to-use API - Supports debugging and visualization - Transparent feature and weight analysis	- Less comprehensive than SHAP/LIME - Limited deep-learning support	- Education, prototyping, model debugging
Alibi	Model-agnostic library	Sparsity	Time-to-understanding	- Covers counterfactual,	- Requires technical expertise	- Enterprise ML pipelines

(*continued on next page*)

Table 2 (*continued*)

XAI tool	Type	Core metrics	Supporting / human-centered metrics	Strengths	Weaknesses / limitations	Possible domains
		Explanation length Coverage Rule length (Anchors)	User trust score	anchors, adversarial detection - Strong support for fairness evaluation	- Less widely documented	- Sensitive domains requiring fairness
IBM AIX360	Comprehensive XAI framework	Explanation length / complexity Rule length Simulatability Coverage	User trust score	- Extensive algorithms + documentation - Open-source, enterprise-ready - Supports datasets + model explainability	- Large and complex ecosystem - Potential steep learning curve	- Regulated industries (finance, healthcare) - Enterprises needing governance support
Microsoft InterpretML	Comprehensive XAI framework	Simulatability Explanation length Visualization clarity Coverage	Time-to-understanding	- Supports interpretable models (EBMs) - Unified dashboard for explanations - Strong community support	- Less tailored for deep learning - Integration mainly in Python ecosystem	- Healthcare, HR, education - Systems needing interpretable boosting models
Google what-if tool (WIT)	Visual interface	Visualization clarity Coverage	Task performance improvement User trust score	- No-code/low-code exploration - Intuitive fairness and performance evaluation - Highly accessible	- Limited support for large-scale or custom DL architectures - Requires TensorBoard integration	- Ethical AI reviews - Education & training - Exploratory fairness analysis
DARPA XAI program	Research & evaluation framework	User trust score Task performance improvement Time-to-understanding	Explanation satisfaction Mental model accuracy	- Integrates cognitive psychology and human reasoning - Supports interpretable ML + post-hoc methods - Strong evaluation criteria (Fidelity, Completeness, Simplicity)	- Research-oriented; less plug-and-play - High complexity, diverse methodologies	- Defense, critical infrastructures - Human-AI collaboration research
Microsoft prediction–decision–recommendation (PDR) framework	AI governance & workflow framework	Task performance improvement Time-to-understanding User trust score	Visualization clarity	- Aligns predictions with human values - Designed for enterprise-scale recommender systems - Supports qualitative + quantitative metrics	- Tailored to recommendation ecosystems - Limited uptake outside Microsoft platforms	- Recommender systems (retail, media) - Decision-support platforms
IEEE P7003 algorithmic bias standard	Ethical & technical standard	Coverage Documentation completeness	User trust score (organizational)	- Provides actionable framework for bias mitigation - Widely recognized ethical standard - Supports documentation + governance	- Not a technical tool—needs developer interpretation - Compliance may require significant restructuring	- Public sector AI - HR and recruitment systems - Safety-critical decision systems
Fairlearn	Fairness assessment & mitigation library	Coverage Visualization clarity	User trust score	- Provides disparity metrics - Offers mitigation algorithms - Integrates with common ML pipelines	- Requires demographic data - Does not explain models—focuses on fairness only	- Credit scoring, insurance, hiring - Any domain requiring fairness constraints
Testing with concept activation vectors (TCAV) (Implemented in Captum)	Concept-based explainability	Simulatability (concept-level) Explanation length	Time-to-understanding User trust score	- Explains models using human-understandable concepts - Helps detect stereotype-driven patterns	- Requires well-defined concepts - Limited to deep models with embeddings	- Computer vision - Medical imaging - NLP conceptual bias detection

(*continued on next page*)

Table 2 (*continued*)

XAI tool	Type	Core metrics	Supporting / human-centered metrics	Strengths	Weaknesses / limitations	Possible domains
Model monitoring for drift (concept drift, covariate shift)	Governance & operational process	Sparsity (concept selection) Coverage Visualization clarity	Task performance improvement (operational) Time-to-understanding (alerts)	- Essential for long-term reliability - Supports proactive correction - Aligns with regulatory expectations	- Requires continuous data pipelines - Resource-intensive in large-scale systems	- Finance (risk models) - Healthcare (diagnostics) - Dynamic environments (e-commerce)

XAI systems require ongoing governance and maintenance once they have been deployed. This includes version control, retraining protocols that are guided by explainability objectives, as well as user feedback mechanisms that support continuous learning and improvement outcomes. The extant literature clearly distinguishes between ante-hoc and post-hoc approaches to explainability (Confalonieri et al., 2021; Kenny et al., 2021; Minh et al., 2022). Inherently interpretable models such as linear regression, rule-based systems, decision trees, GAMs and Bayesian models are transparent by design. Such models enable users to directly understand their modus operandi, operational logic and decision-making processes. By contrast, black-box models, including deep neural networks, necessitate post-hoc interpretability methods. Techniques such as SHAP and LIME provide feature-attribution and local explanations (Adak et al., 2022; Barnard et al., 2022; Li, 2022; Shams Amiri et al., 2021; Yang et al., 2024; Zhang et al., 2022), while counterfactual reasoning, fairness audits and human-in-the-loop (HITL) approaches are increasingly employed to enhance transparency, accountability and equity in high-stakes contexts (Holzinger, 2021; Mosqueira-Rey et al., 2023; Olson et al., 2021).

This review confirms that SHAP offers model-agnostic explanations by quantifying the contribution of individual features to model outputs, whereas LIME explains specific predictions by locally approximating complex models with interpretable surrogates. In addition, other open-source tools (e.g. ELI5, Alibi) and commercial platforms (e.g. IBM AIX360, Microsoft InterpretML, Google's What-If Tool) have expanded the XAI ecosystem. Methodological approaches such as counterfactual explanations further support understanding by exploring "what-if" scenarios, while ongoing fairness audits evaluate model behaviors across demographic groups, to identify and mitigate bias. Human-in-the-loop (HITL) approaches complement these techniques by embedding human oversight throughout the AI lifecycle, thereby strengthening contextual accuracy and accountability.

Additionally, several institutions initiatives have led to the formalization of XAI assessment and evaluation standards. For instance, the DARPA XAI Program features quantitative metrics (such as fidelity, completeness, simplicity, robustness and performance), as well as

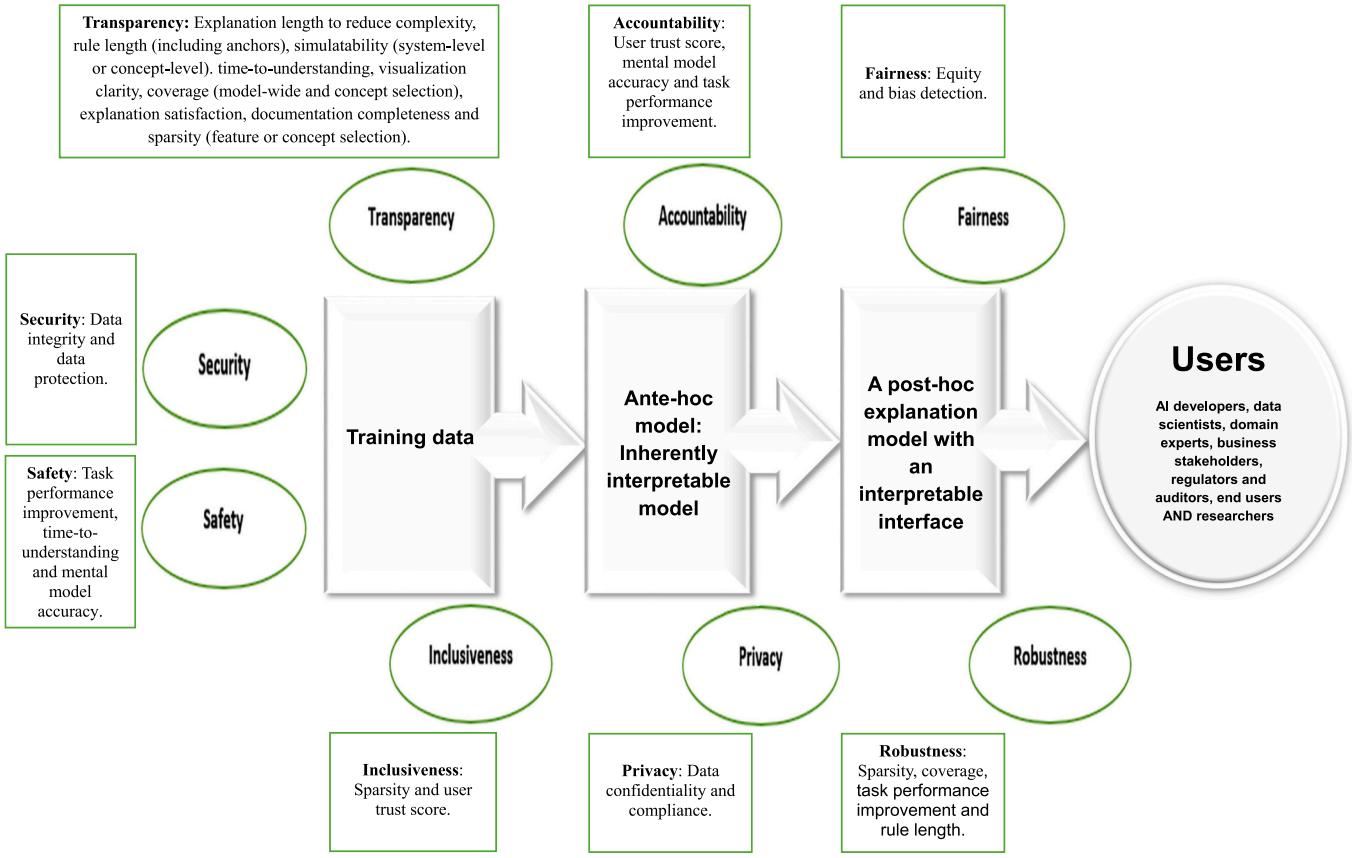

Fig. 2. A user-centric explainable artificial intelligence (XAI) framework for black box models.

qualitative ones (including human-centered evaluations that examine perceived usefulness, trust, satisfaction and task effectiveness). Yet, despite these advances, many existing XAI approaches remain technique-specific, as they exclusively focus on post-hoc explanations, fairness audits or concept-based methods, often resulting in fragmented evaluation practices.

Against this backdrop, this research puts forward an easy-to-understand, user-centric XAI framework for black-box models. This conceptual framework raises awareness on human-centered evaluation metrics and integrates them as a unifying analytical lens across the AI lifecycle (rather than assessing explainability in isolation). It explicitly links data practices, model design choices and explanation interfaces to measurable user outcomes, as illustrated in Fig. 2.

Firstly, this user-centric XAI framework emphasizes transparent, inclusive and secure training data as a foundation for explainability and trust. While governance-oriented tools and standards (e.g. Fairlearn, IEEE P7003) primarily support compliance and bias detection, this model suggests that inclusiveness, transparency, safety and security metrics during the training phase by ensure that the models are developed in a manner that is fair, interpretable, robust and trustworthy.

The inclusiveness metrics help detect and mitigate biases in training data and model behavior, thereby promoting fairness. They ensure objective and consistent performance of AI systems across diverse user groups. Hence, they lead to explanations that are meaningful and relevant to all stakeholders. The transparency metrics are meant to evaluate how clearly the model's internal decision-making processes can be understood by their users. During training, these metrics guide the development of models that produce interpretable and accessible explanations, in order to improve user comprehension and trust.

The safety metrics monitor the model's behavior under various conditions, including during unusual, rare or unexpected situations (a.k.a. edge cases) that challenge the system's robustness, to prevent harmful or unintended outcomes. The integration of safety considerations in training enhance the systems' reliability aspects, as they ensure that explanations reflect typical contexts as well as exceptional (or even risky) scenarios. Similarly, the security metrics assess vulnerabilities to adversarial attacks or data manipulation. The inclusion of security metrics in training, models become more robust, and their explanations would enhance confidence levels and reduce potential risks, thereby fostering greater user assurance.

Secondly, the framework incorporates an accountable ante-hoc model layer grounded in inherently interpretable models. Clearly, it is consistent with decision trees and rule-based systems, as this layer prioritizes sparsity, simulatability and explanation conciseness. It facilitates quick understanding and mental simulation of decisions. In doing so, it and accountability beyond what post-hoc methods alone can achieve. The accountability metrics reinforces predictability and can strengthen the trustworthiness and governance of AI systems by: (i) evaluating whether the model's decision logic can be audited and traced, thereby ensuring each prediction can be explained and justified to stakeholders; (ii) ensuring compliance with ethical and legal standards; (iii) assessing stakeholder understanding and acceptance; and, (iv) facilitating error and bias detection.

There is scope for practitioners to incorporate accountability metrics, if they want their inherently interpretable models to become more auditable, responsible and trustworthy. At the same time, they can enhance the value of ante-hoc explainability by adopting privacy metrics that safeguard sensitive information throughout the interpretability process. Though inherently interpretable models are transparent by design, the privacy metrics would ensure that this transparency does not compromise sensitive data by: (i) measuring risk of sensitive (personal) information exposure; (ii) enforcing data minimization principles to ensure that the model uses only the indispensable data to reduce privacy risks; (iii) Balancing interpretability and data protection (e.g. through anonymization techniques) to maintain explainability while respecting privacy constraints; and, (iv) Supporting compliance with data protection regulations (E.g. by complying with GDPR or other relevant privacy laws).

Thirdly, this framework integrates fair and robust post-hoc explanations with interpretable user interfaces. While tools such as SHAP, LIME, Alibi, and TCAV are commonly evaluated using metrics such as sparsity, complexity, and visualization clarity, this framework extends their application by explicitly prioritizing trust calibration and task performance improvement, particularly when AI systems are employed for decision support in human-in-the-loop (HITL) settings. This emphasis aligns with human-centered evaluation principles advocated in initiatives such as DARPA XAI and Microsoft's Prediction–Decision–Recommendation (PDR) framework.

Post-hoc explanation methods are applied after a black-box model has been trained and has generated predictions (e.g., SHAP, LIME, and counterfactual explanations). While fairness metrics (e.g., demographic parity, equalized odds, and disparate impact) quantify whether the model's decisions are biased or discriminatory across different demographic groups, robustness metrics assess stability under perturbations, including both predictive robustness (i.e., stability of model outputs) and explanation robustness (i.e., consistency of explanations under slight input variations). In this context, robustness may refer both to the stability of the model's predictions and to the consistency of the generated explanations under slight variations in the input data.

The fairness and robustness metrics build user trust and enhance XAI in post-hoc settings as they reveal biases, validate explanation reliability (as explanations are not expected to significantly change if they are meeting and exceeding their robust performance metrics), guide explanation refinement (by monitoring fairness and robustness metrics, developers can fine tune post-hoc methods to produce explanations that are accurate and fairly representative of the model's decision logic), improve interface transparency, and support regulatory compliance as well as ethical standards that foster increased transparency and accountability of AI systems.

Overall, this conceptual framework offers a coherent, user-oriented benchmark for assessing explainability across data, models, and interfaces, thereby extending existing XAI frameworks developed by technology firms and standards bodies. It implies that ante-hoc (inherently interpretable) models can inform and calibrate post-hoc explanation methods and their associated interfaces. Ante-hoc models may serve as interpretable baselines against which the fidelity and consistency of post-hoc explanations from black-box models are assessed. Therefore, the integration of ante-hoc and black-box models can support the development of more trustworthy systems, particularly by enabling interpretable interfaces to be trained or tested against transparent model logic before deployment in more complex settings. Accordingly, this framework positions ante-hoc models as an intermediary layer between training data and post-hoc explanations. This enables explanation methods and interfaces to be validated against interpretable model logic before being applied to complex black-box systems.

4. Conclusions

This research synthesizes key contributions in XAI to underline its essential role in promoting responsible governance in the research, development and maintenance of machine learning systems. It discusses about XAI tools, describes their metrics, identifies their strengths as well as their weaknesses / limitations. Moreover, it reports their possible domains. It addresses ethical concerns related to black-box models. Hence, it emphasizes the need for documentation practices that establish the normative and technical baselines for accountability, upon which performance tracking and continuous monitoring are built. One has to consider that robust drift detection and fairness auditing are dependent on these baselines and operate iteratively throughout deployment in order to maintain reliable, transparent and equitable XAI systems.

This contribution's user-centric XAI framework with its interpretable interfaces that bridge technical innovation and stakeholder ethics are

intended to foster responsible AI and ensure that ML models remain interpretable, trustworthy and compliant with ethical and legal standards like GDPR and the EU AI Act, throughout their lifecycle.

4.1. Theoretical implications

This research adds value to the extant academic literature focused on XAI (Buçinca et al., 2021; Li, 2022; Minh et al., 2022; Shin, 2021; Yeo et al., 2025). It clarifies key notions and explains the meanings of different terms related to model interpretability, data drift, concept drift and fairness. Moreover, it clarifies how practitioners can build and maintain trustworthy AI systems. It clearly indicates that interpretability is a crucial mechanism for fostering user trust, not just through technical explanations, but also by adhering to clear governance structures and established communication channels (Baptista et al., 2022; Graziani et al., 2023; Iban, 2022; Yang et al., 2024). This reasoning aligns with emerging theories related to human-computer interaction and to technology adoption frameworks drawn from social sciences literature (Camilleri and Troise, 2023), that highlight the importance of transparency and accountability in building user confidence in complex systems.

This research builds on the foundations of established theoretical underpinnings by integrating explainable AI within broader models of technology acceptance, trust and socio-technical dynamics (Camilleri, 2024). For example, some elements of this contribution's conceptual framework are related to the Technology Acceptance Model (TAM) key constructs including to perceived usefulness and perceived ease of use, as these factors clearly align with XAI's goals of enhancing transparency and interpretability of ML models to foster user adoption. The framework also draws on Trust in Automation theories, particularly where they highlight the rationale for the development of explainable AI systems, to enhance user trust, and to prevent their misuse or disuse. In a similar vein, some commentators argue that XAI literature is grounded in the Socio-Technical Systems (STS) theory (Appelbaum, 1997; Trist, 1981). They contend that this theory provides a holistic lens by emphasizing the interplay between technological artifacts and social contexts, thereby reinforcing the need for inclusive, ethical and transparent AI design. Other colleagues maintain that XAI literature is rooted in Responsible Research and Innovation (RRI) frameworks as they raise awareness about anticipatory governance, stakeholder engagement and ethical reflexivity (EU, 2014; Stahl and Wright, 2018), all of which are operationalized through user-centric and transparent approaches. Together, these models serve as a theoretical basis for this study's conceptual framework, as they bridge technical, human, ethical and regulatory dimensions to support trustworthy AI ecosystems.

This timely contribution promotes transparent and fair forms of AI knowledge generation, as the reasoning behind ML decisions and predictions ought to be continuously scrutinized and validated. It puts forward a comprehensive framework that synthesizes key dimensions of XAI into a cohesive model. It reports how, why, where and when explainability is evolving within generative AI systems. Generally, by linking design choices to measurable user outcomes across the AI lifecycle. Unlike prior models that are narrowly focused on interpretability techniques, this framework integrates lifecycle governance with human-centered evaluation metrics. It supports the practical implementation of responsible AI principles. By doing so, it advances theoretical understanding while offering actionable guidance for developers, policymakers and stakeholders committed to trustworthy AI.

In sum, it provides a comprehensive explanation of XAI systems for the benefit of their users including AI developers, data scientists, domain experts, business stakeholders, regulators and auditors, end users as well as academic researchers, among others. It enables them to better understand the modus operandi of deep neural networks and complex learning models. It promotes post-hoc explanation techniques and methods that provide explanations for the decisions made by machine learning models after they have been trained (Confalonieri et al., 2021; Kenny et al., 2021; Mosqueira-Rey et al., 2023; Panigutti et al., 2021). This is particularly important for opaque black box models, ensemble methods or support vector machines, which offer high predictive accuracy but are not clear enough on how they arrive at specific outputs. It identifies XAI tools that can help practitioners assess the validity and reliability of ML models.

This research emphasizes the dynamic challenges of AI deployment. It makes reference to model drift and to data distribution shifts, as they can have a negative impact on the reliability and fairness of explanations over time (Hinder et al., 2023; Rahmani et al., 2023). This perspective moves beyond static evaluations of XAI. It highlights the need for continuous monitoring and adaptation of AI models. It considers the needs and challenges faced not only by AI developers but also by system administrators and non-expert users. It recognizes that effective XAI must cater to diverse levels of technical understanding and operational requirements.

This article also offers novel, integrated and up-to-date syntheses of both academic research as well as practitioner-oriented tools and frameworks. It bridges the gap between theoretical advancements and their real-world applications across the entire AI lifecycle. It refers to technical aspects including XAI specific tools and techniques, data monitoring, fairness assurance and stakeholder engagement, thereby providing a timely and holistic view of the current XAI landscape.

4.2. Practical iplications

This research offers guidance for a wide range of stakeholders involved in the development, deployment and governance of AI systems. It provides actionable insights for developers and system administrators for implementing XAI. It describes specific tools (e.g. SHAP, LIME, ELI5) and platforms that offer concrete entry points for integrating interpretability into their workflows. This article highlights a comparison matrix of leading XAI tools. It outlines their key metrics, strengths, limitations and domain suitability to support informed managerial decision-making.

Additionally, it proposes a user-centric XAI framework tailored for black-box models. This framework offers practical guidance on aligning explainability techniques with organizational capabilities, stakeholder expectations, and contextual constraints. The novel framework provides a tangible structure that embeds responsible AI practices from the initial design phase through ongoing monitoring and updates. It is intended to support practitioners in the development of more robust, reliable and trustworthy AI applications. Its recommendations for the integration of interpretability, regular bias monitoring and fairness auditing (through standardized reporting frameworks, such as model cards and data sheets, combined with automated drift detection tools) can inform policy makers as well as practitioners to advance XAI systems. Hence, the development of internal policies, quasi-substantive rules and workflows are intended to advance responsible AI development and deployment. This may ultimately lead to virtuous outcomes that are intended to foster a culture of ethical AI innovation that enhances public trust and understanding of XAI systems, leading to increased user adoption in different domains.

4.3. Limitations and future research directions

Despite its contributions, this study also has its inherent limitations. The systematic review involved the analysis of recent, high-impact academic publications focused on “explainable artificial intelligence” or “explainable AI” or “XAI”. This selection approach, while ensuring relevance and quality, introduces the risk of citation bias, where frequently cited or well-known studies receive disproportionate attention, potentially overshadowing emerging, less-cited, or interdisciplinary work. Consequently, some innovative advancements or niche applications in XAI may not have been fully captured. Additionally, the quickly evolving nature of the field means new developments could have

emerged after the review period. Furthermore, the evaluation of XAI tools and frameworks relied on publicly available information and academic studies, which often lack empirical depth or comprehensive real-world validation, thereby limiting the scope for fully assessing practical performance and impact of interpretable models.

Future research can address these limitations and explore plausible areas of study related to XAI. For example, there is scope for conducting longitudinal studies to examine the long-term impact of XAI adoption on system performance, user trust and on the fairness of AI outputs in real-world scenarios. Moreover, other research is required to develop standardized metrics that can evaluate the "quality" of explanations and their effectiveness for different user groups hailing from diverse contexts. Perhaps, prospective researchers can build on this seminal article by promoting the integration of XAI techniques with other responsible AI governance frameworks, such as privacy-secure AI methodologies, robust AI, as well as inclusive, bias-free AI systems in the near future. In addition, they may analyze human-computer interaction aspects of XAI, including how different types of explanations are perceived and understood by diverse stakeholders. It is imperative that developers design effective and interpretable user-centric XAI solutions. Further research in these fields of study will contribute to the continued advancement and to the responsible adoption of explainable AI, as shown in Table 3.

Table 3
Future research directions related to explainable AI (XAI).

Future research area	Rationale	Potential impact
Context-specific XAI	To investigate user backgrounds, domain knowledge of XAI and cultural contexts.	Increases usability and accessibility of XAI systems.
Human-computer interaction (HCI) in XAI	To explore how different stakeholders perceive, interpret and interact with different types of AI explanations.	Improves the design of user-centric and interpretable XAI solutions.
Focus on niche and emerging XAI applications	To examine XAI applications in specialized domains (e.g., healthcare, finance, autonomous systems).	Expands XAI applicability and domain-specific innovations.
Integration of XAI with responsible AI governance frameworks	To better understand how XAI can be associated with privacy-preserving, robust and bias-free AI methodologies, to advance holistic AI governance frameworks.	Promotes trustworthy, fair, and secure XAI deployment.
Empirical validation of XAI tools and frameworks	In-depth and broad empirical studies will shed light on the effectiveness of current XAI tools in real-world applications.	Bridges the gap between theoretical models and practical uses of XAI.
Longitudinal studies on XAI adoption	To analyze the long-term effects of XAI on system performance, user trust and fairness in real-world contexts.	Advances knowledge on sustained benefits and risks of XAI use.
Ethical and social implications of XAI	To demonstrate the societal impacts, ethical challenges and policy considerations arising from XAI adoption.	Guides responsible AI governance deployment that respects societal norms.
Development of standardized evaluation metrics	To create standardized, reliable metrics that can assess XAI quality and its effectiveness across diverse users.	Enables consistent benchmarking and comparison of XAI tools.

CRediT authorship contribution statement

Mark Anthony Camilleri: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization.

Author's statements

The author declares that he does not have any conflicting or competing interests.

The author declares that he did not avail himself from external funding to conduct this research.

The author declares that he did not use generative artificial intelligence to prepare this manuscript.

The corresponding author is accountable for all aspects of work in this manuscript, including: (1) the conception and design of the study, or acquisition of data, or analysis and interpretation of data; (2) Drafting the article or revising it critically for important intellectual content; AND (3) Final approval of the version to be submitted.

Declaration of competing interest

The author declares he has no conflicts of interest.

Appendix A. Key concepts in explainable artificial intelligence research.

XAI key term	Description
Accountability	Accountability ensures that individuals or organizations can be held responsible for the outcomes and impacts of AI systems, especially in critical applications, where errors or biases could have significant consequences. Individuals and organizations ought to be supported by clear, interpretable explanations that enable oversight and compliance with ethical or regulatory standards.
Artificial Intelligence (AI)	AI is a broad field in computer science focused on creating machines that can perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception and decision-making.
Black box / Black-box model	The black box (model) refers to the opaque nature of various AI models like deep neural networks' decision-making processes (as users including their developers may not be in a position to understand their modus operandi). While such models can usually achieve high accuracy, they may not be transparent about how their data processing works and on how they produce a specific output.
Counterfactual explanations	Counterfactual explanations are a type of model-agnostic explanation technique used in interpretable and explainable AI (XAI). They describe how an input instance would need to be altered minimally for a machine learning model to yield a different (usually a desired) outcome.
Decision making	Decision-making in the context of AI refers to the process where an AI system uses computational techniques to analyze data, identify patterns, and determine optimal courses of action or choices from a set of alternatives. Unlike human decision-making, which can rely on intuition, experience, or emotion, AI decision-making is data-driven and is based on algorithms.
Decision support systems (DSS)	DSS are applications that analyze data and provide valuable insights. They are designed to assist humans in making informed choices. In XAI contexts, DSS integrate explainability into such systems, and transform them from “black boxes” into transparent tools that users can understand and trust, especially in sensitive domains like healthcare, among others.
Deep Learning (DL)	DL is a subset of machine learning that focuses on utilizing multilayered (deep) neural networks to learn patterns and representations directly from raw data, to discover intricate features and perform tasks such as classification, regression and representation learning.
Evaluation metrics	The evaluation metrics relate to how AI and XAI systems are assessed and measured. They enable practitioners to objectively evaluate their effectiveness as well as the quality of explanations generated by AI systems. While AI models are typically evaluated on their predictive performance (e.g., in terms of their accuracy), XAI evaluation metrics go beyond this to measure how well explanations help users understand, trust and interact with AI systems. In this case, evaluation metrics may include human-centered metrics (e.g. the users' trust and satisfaction levels vis-à-vis XAI) as well as quantitative metrics (like measuring the model's accuracy and comprehensiveness).
Explainable Artificial Intelligence / Explainable AI (XAI)	XAI explores methods that provide humans with the ability and intellectual oversight to understand AI outputs. The rationale behind XAI is to increase the interpretability and transparency of AI decisions, actions and predictions. In other words, XAI is intended to answer the “why” and “how” behind AI’ systems, as they often function as blackboxes.
Feature attribution	Feature attribution refers to the process of quantifying the contribution or importance of each input feature in a machine learning model's prediction. It helps explain how much each feature influences a particular decision made by the model. This is especially valuable in interpretable machine learning and explainable AI (XAI), as the understanding why an AI model is advancing a certain prediction is as important as the prediction itself.
Human-AI interaction (HAII) / Human-computer interaction (HCI)	HAII and HCI concepts and their variations emphasize the user-centricity aspects of AI. In the context of XAI, both notions suggest that humans are more likely to engage, communicate and collaborate with intuitive and explainable AI interfaces.
Human-in-the-Loop (HITL)	HITL approaches refer to systems or processes in AI and ML where human judgement and intervention are actively integrated into the decision-making loop. This involvement can occur at various stages, through data collection, labeling and annotation (often with human input), data preprocessing and curation, model training, model evaluation and validation (with human oversight, especially in high-stakes domains), model deployment as well as during monitoring and maintenance phases. The underlying goal of HITL is to combine the strengths of human intuition, contextual understanding and ethical reasoning with the efficiency and scale of automated systems.
Interpretability	Interpretability is related to the degree to which a human can understand internal mechanics, in terms of the cause-effect relationships of its decision-making processes of XAI models. This construct suggests that users tend to interact with transparent and trustworthy XAI technologies because they can facilitate the interpretation of their outputs.
Local Interpretable Model-agnostic Explanations (LIME)	LIME is a technique that explains individual predictions by approximating a complex model (e.g. in a localized setting), with an interpretable one, such as a linear model. LIME highlights which features could influence a specific decision (by perturbing input data) to observe how predictions change, thereby making black-box models more understandable to users without requiring access to their internal structures.
Machine learning	ML is a field in AI concerned with the development and study of algorithms that can identify patterns within the data. This allows them to learn from them and to make decisions as well as predictions. Such systems can perform tasks without explicit instructions and could improve their performance over time, as they are exposed to more data.
Mental models / Shared mental models	Shared mental models refer to the mutual understanding and common representation of knowledge between humans and AI agents regarding their respective roles, capabilities and their task at hand. Essentially, they refer to the extent to which there is a shared understanding of how the AI system operates and how it aligns with the overall task.
Neural networks (models)	Neural networks are complex machine learning architectures with interconnected “layers” used to learn patterns from data, to perform specific tasks like predictions or classifications. The role of XAI is to provide explanations about how such opaque networks/models work. It clarifies how inputs influence outputs and reveals what the AI model has learned.
Perturbation analysis	Perturbation analysis involves systematically altering (perturbing) one or more features of the input data and observing how the model's output changes.
Post-hoc explanations	Post-hoc explanations are retrospective interpretability techniques that are used to explain the predictions of already trained machine learning models after they have made a decision. Post-hoc explanations are generated after model training and are not part of the original learning process. They aim to interpret how or why a model made a specific decision, without altering the model itself.
SHapley Additive exPlanations (SHAP)	SHAP is a method based on Shapley values from cooperative game theory. It is used to explain the output of machine learning models. Basically, SHAP offers consistent and theoretically grounded insights into how individual XAI features contribute to its model decisions, by assigning each feature an “importance value” for a specific prediction. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is.
Transparency	Transparency refers to the clarity and understandability of an AI system's internal workings and decision-making processes. Hence, it allows humans to learn how AI systems process data and make decisions.
Trust	Trust refers to the users' confidence levels they place on XAI systems' decisions. The individuals' willingness to avail themselves of XAI technologies relies on their reliability in terms of clarity, consistency and usefulness of their explanations. XAI aims to foster

(*continued on next page*)

(*continued*)

XAI key term	Description
	appropriate levels of trust by helping users to better understand how and why AI models make certain decisions, predictions or generate outcomes.
User behavior / User study	User behavior focuses on how individuals interact with, perceive, and respond to the explanations provided by AI systems. The persons' cognitive processes, trust, decision-making and reliance on AI systems can influence their engagement levels with XAI technologies.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.

Adak, A., Pradhan, B., Shukla, N., 2022. Sentiment analysis of customer reviews of food delivery services using deep learning and explainable artificial intelligence: systematic review. Foods 11 (10), 1500.

AI Act, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union Commission, Brussels, Belgium. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Andrews, R.W., Lilly, J.M., Srivastava, D., Feigh, K.M., 2023. The role of shared mental models in human-AI teams: a theoretical review. Theor. Issues Ergon. Sci. 24 (2), 129–175.

Appelbaum, S.H., 1997. Socio-technical systems theory: an intervention strategy for organizational development. Manag. Decis. 35 (6), 452–463.

Askr, H., Elgeldawi, E., Aboul Ella, H., Elshaier, Y.A.M.M., Gomaa, M.M., Hassanien, A. E., 2023. Deep learning in drug discovery: an integrative review and future challenges. Artif. Intell. Rev. 56 (7), 5975–6037.

Bao, H., Liu, W., Dai, Z., 2025. Artificial intelligence vs. public administrators: public trust, efficiency, and tolerance for errors. Technol. Forecast. Soc. Chang. 215, 124102.

Baptista, M.L., Goebel, K., Henriques, E.M.P., 2022. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. Artif. Intell. 306, 103667.

Barnard, P., Marchetti, N., Dasilva, L.A., 2022. Robust network intrusion detection through explainable artificial intelligence (XAI). IEEE Networking Letters 4 (3), 167–171.

Bauer, K., von Zahn, M., Hinz, O., 2023. Expl(AI)ned: the impact of explainable artificial intelligence on users' information processing. Inf. Syst. Res. 34 (4), 1582–1602.

Buçinca, Z., Malaya, M.B., Gajos, K.Z., 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW1), 1–21.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N., 2023. A survey on XAI and natural language explanations. Inf. Process. Manag. 60 (1), 103111.

Camilleri, M.A., 2023. Artificial intelligence governance: ethical considerations and implications for social responsibility. Expert. Syst. 41 (7), e13406.

Camilleri, M.A., 2024. Factors affecting performance expectancy and intentions to use ChatGPT: using SmartPLS to advance an information technology acceptance framework. Technol. Forecast. Soc. Chang. 201. https://doi.org/10.1016/j.techfore.2024.123247.

Camilleri, M.A., Troise, C., 2023. Live support by chatbots with artificial intelligence: a future research agenda. Serv. Bus. 17 (1), 61–80.

Chen, V., Liao, Q.V., Wortman Vaughan, J., Bansal, G., 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. Proceedings of the ACM on Human-Computer Interaction 7 (CSCW2), 1–32.

Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., Melacci, S., 2023. Logic explained networks. Artif. Intell. 314, 103822.

Conati, C., Barral, O., Putnam, V., Rieger, L., 2021. Toward personalized XAI: a case study in intelligent tutoring systems. Artif. Intell. 298, 103503.

Confalonieri, R., Weyde, T., Besold, T.R., del Prado Martín, F.M., 2021. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. Artif. Intell. 296, 103471.

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., Cruz, F., 2021. Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif. Intell. 299, 103525.

De Bruijn, H., Warnier, M., Janssen, M., 2022. The perils and pitfalls of explainable AI: strategies for explaining algorithmic decision-making. Gov. Inf. Q. 39 (2), 101666.

Deepa, R., Sekar, S., Malik, A., Kumar, J., Attri, R., 2024. Impact of AI-focussed technologies on social and technical competencies for HR managers–a systematic review and research agenda. Technol. Forecast. Soc. Chang. 202, 123301.

Di Martino, F., Delmastro, F., 2023. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artif. Intell. Rev. 56 (6), 5261–5315.

Dikmen, M., Burns, C., 2022. The effects of domain knowledge on trust in explainable AI and task performance: a case of peer-to-peer lending. Int. J. Hum.-Comput. Stud. 162, 102792.

Dwivedi, Y.K., Sharma, A., Rana, N.P., Giannakis, M., Goel, P., Dutot, V., 2023. Evolution of artificial intelligence research in technological forecasting and social change: research topics, trends, and future directions. Technol. Forecast. Soc. Chang. 192, 122579.

EU, 2014. Rome Declaration on Responsible Research and Innovation in Europe. European Commission, Brussels, Belgium. https://digital-strategy.ec.europa.eu/en/library/rome-declaration-responsible-research-and-innovation-europe.

Fan, M., Mo, Z., Zhao, Q., Liang, Z., 2024. Innovative insights into knowledge-driven financial distress prediction: a comprehensive XAI approach. J. Knowl. Econ. 15 (3), 12554–12595.

Garcez, A.D.A., Lamb, L.C., 2023. Neurosymbolic AI: the 3rd wave. Artif. Intell. Rev. 56 (11), 12387–12406.

Ghaffarian, S., Taghikhah, F.R., Maier, H.R., 2023. Explainable artificial intelligence in disaster risk management: achievements and prospective futures. International Journal of Disaster Risk Reduction 98, 104123.

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J.P., Yordanova, K., Vered, M., Müller, H., 2023. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. Artif. Intell. Rev. 56 (4), 3473–3504.

Guba, E.G., Lincoln, Y.S., 1985. Naturalistic Inquiry. Sage Publishing, Beverly Hills, CA, United States of America.

Guidotti, R., 2021. Evaluating local explanation methods on ground truth. Artif. Intell. 291, 103428.

Guleria, P., Sood, M., 2023. Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. Educ. Inf. Technol. 28 (1), 1081–1116.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z., 2019. XAI—explainable artificial intelligence. Sci. Robot. 4 (37), eaay7120.

Herm, L.V., Heinrich, K., Wanner, J., Janiesch, C., 2023. Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. Int. J. Inf. Manag. 69, 102538.

Hinder, F., Vaquet, V., Brinkrolf, J., Hammer, B., 2023. Model-based explanations of concept drift. Neurocomputing 555, 126640.

Holzinger, A., 2021. Explainable AI and multi-modal causability in medicine. i-com 19 (3), 171–179.

Iban, M.C., 2022. An explainable model for the mass appraisal of residences: the application of tree-based machine learning algorithms and interpretation of value determinants. Habitat Int. 128, 102660.

Jang, Y., Choi, S., Jung, H., Kim, H., 2022. Practical early prediction of students' performance using machine learning and eXplainable AI. Educ. Inf. Technol. 27 (9), 12855–12889.

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., Kuk, G., 2022. Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. Soc. Sci. Comput. Rev. 40 (2), 478–493.

Jiménez-Luna, J., Skalic, M., Weskamp, N., Schneider, G., 2021. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. J. Chem. Inf. Model. 61 (3), 1083–1094.

Johnson, M., Albizri, A., Harfouche, A., Fosso-Wamba, S., 2022. Integrating human knowledge into artificial intelligence for complex and ill-structured problems: informed artificial intelligence. Int. J. Inf. Manag. 64, 102479.

Joung, J., Kim, H., 2023. Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. Int. J. Inf. Manag. 70, 102641.

Kenny, E.M., Ford, C., Quinn, M., Keane, M.T., 2021. Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif. Intell. 294, 103459.

Khosravi, H., Shum, S.B., Chen, G., Conati, C., Tsai, Y.S., Kay, J., Gašević, D., 2022. Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence 3, 100074.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Baum, K., 2021. What do we want from explainable artificial intelligence (XAI)?–a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif. Intell. 296, 103473.

Li, Z., 2022. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. Comput. Environ. Urban. Syst. 96, 101845.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable ai: a review of machine learning interpretability methods. Entropy 23 (1), 18.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2 (1), 56–67.

Malach, A., Wudali, P.N., Momiyama, S., Furukawa, J., Araki, T., Elovici, Y., Shabtai, A., 2025. CyberShapley: explanation, prioritization, and triage of cybersecurity alerts using informative graph representation. Comput. Secur. 150, 104270.

Meske, C., Bunde, E., Schneider, J., Gersch, M., 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. Inf. Syst. Manag. 39 (1), 53–63.
Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N., 2022. Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. 1–66.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339. https://doi.org/10.1136/bmj.b2535.
Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, Á., 2023. Human-in-the-loop machine learning: a state of the art. Artif. Intell. Rev. 56 (4), 3005–3054.
Naiseh, M., Al-Thani, D., Jiang, N., Ali, R., 2023. How the different explanation classes impact trust calibration: the case of clinical decision support systems. International Journal of Human-Computer Studies 169, 102941.
Olson, M.L., Khanna, R., Neal, L., Li, F., Wong, W.K., 2021. Counterfactual state explanations for reinforcement learning agents via generative deep learning. Artif. Intell. 295, 103455.
Owen, R., Von Schomberg, R., Macnaghten, P., 2021. An unfinished journey? Reflections on a decade of responsible research and innovation. Journal of Responsible Innovation 8 (2), 217–233.
Oyebode, O., Fowles, J., Steeves, D., Orji, R., 2023. Machine learning techniques in adaptive and personalized systems for health and wellness. Int. J. Hum. Comput. Interact. 39 (9), 1938–1962.
Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., Pedreschi, D., 2021. FairLens: auditing black-box clinical decision support systems. Inf. Process. Manag. 58 (5), 102657.
Rahman, M.A., Hossain, M.S., Showail, A.J., Alrajeh, N.A., Alhamid, M.F., 2021a. A secure, private, and explainable IoHT framework to support sustainable health monitoring in a smart city. Sustain. Cities Soc. 72, 103083.
Rahman, M.A., Hossain, M.S., Showail, A.J., Alrajeh, N.A., Alhamid, M.F., 2021b. A secure, private, and explainable IoHT framework to support sustainable health monitoring in a smart city. Sustain. Cities Soc. 72, 103083.
Rahmani, K., Thapa, R., Tsou, P., Chetty, S.C., Barnes, G., Lam, C., Tso, C.F., 2023. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. Int. J. Med. Inform. 173, 104930.
Salem, A.H., Azzam, S.M., Emam, O.E., Abohany, A.A., 2024. Advancing cybersecurity: a comprehensive review of AI-driven detection techniques. J. Big Data 11 (1), 105.
Sanneman, L., Shah, J.A., 2022. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. Int. J. Hum.-Comput. Interact. 38 (18–20), 1772–1788.
Schoonderwoerd, T.A., Jorritsma, W., Neerincx, M.A., Van Den Bosch, K., 2021. Human-centered XAI: developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies 154, 102684.
Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F., 2021. Glocalx-from local to global explanations of black box ai models. Artif. Intell. 294, 103457.
Shajalal, M., Boden, A., Stevens, G., 2024. ForecastExplainer: explainable household energy demand forecasting by approximating shapley values using DeepLIFT. Technol. Forecast. Soc. Chang. 206, 123588.
Shams Amiri, S., Mottahedi, S., Lee, E.R., Hoque, S., 2021. Peeking inside the black-box: explainable machine learning applied to household transportation energy consumption. Comput. Environ. Urban. Syst. 88, 101647.
Shin, D., 2021. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. International journal of Human-Computer Studies 146, 102551.
Singu, H.B., Chakraborty, D., Troise, C., Camilleri, M.A., Bresciani, S., 2026. Responsible AI for trustworthy tourism: a framework for mitigating ambiguity and anxiety with generative AI. Technol. Forecast. Soc. Chang. 223. https://doi.org/10.1016/j.techfore.2025.124407.
Stahl, B.C., Wright, D., 2018. Ethics and privacy in AI and big data: implementing responsible research and innovation. IEEE Security & Privacy 16 (3), 26–33.
Sun, Z., Song, D., Hero, A., 2023. Minimum-risk recalibration of classifiers. Adv. Neural Inf. Proces. Syst. 36, 69505–69531.
Susnjak, T., Ramaswami, G.S., Mathrani, A., 2022. Learning analytics dashboard: a tool for providing actionable insights to learners. Int. J. Educ. Technol. High. Educ. 19 (1), 12.
Tiddi, I., Schlobach, S., 2022. Knowledge graphs as tools for explainable machine learning: a survey. Artif. Intell. 302, 103627.
Trist, E.L., 1981. The Evolution of Socio-Technical Systems, 2. Ontario Quality of Working Life Centre, Toronto, p. 1981.
van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M., 2021. Evaluating XAI: a comparison of rule-based and example-based explanations. Artif. Intell. 291, 103404.
Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S., Krishna, R., 2023. Explanations can reduce overreliance on ai systems during decision-making. Proceedings of the ACM on Human-Computer Interaction 7 (CSCW1), 1–38.
Wang, L., Wang, D., Tian, F., Peng, Z., Fan, X., Zhang, Z., Wang, H., 2021. Cass: towards building a social-support chatbot for online health community. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW1), 1–31.
Wang, N., Guo, Z., Shang, D., Li, K., 2024. Carbon trading price forecasting in digitalization social change era using an explainable machine learning approach: the case of China as emerging country evidence. Technol. Forecast. Soc. Chang. 200, 123178.
Wysocki, O., Davies, J.K., Vigo, M., Armstrong, A.C., Landers, D., Lee, R., Freitas, A., 2023. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. Artif. Intell. 316, 103839.
Yang, L., Yang, H., Yu, B., Lu, Y., Cui, J., Lin, D., 2024. Exploring non-linear and synergistic effects of green spaces on active travel using crowdsourced data and interpretable machine learning. Travel Behav. Soc. 34, 100673.
Yang, S.J., Ogata, H., Matsui, T., Chen, N.S., 2021. Human-centered artificial intelligence in education: seeing the invisible through the visible. Computers and Education: Artificial Intelligence 2, 100008.
Yang, Z., Zhang, W., Feng, J., 2022. Predicting multiple types of traffic accident severity with explanations: a multi-task deep learning framework. Saf. Sci. 146, 105522.
Yeo, W.J., Van Der Heever, W., Mao, R., Cambria, E., Satapathy, R., Mengaldo, G., 2025. A comprehensive review on financial explainable AI. Artif. Intell. Rev. 58 (6), 1–49.
Yigitcanlar, T., Mehmood, R., Corchado, J.M., 2021. Green artificial intelligence: towards an efficient, sustainable and equitable technology for smart cities and futures. Sustainability 13 (16), 8952.
Zhang, Z., Wu, C., Qu, S., Chen, X., 2022. An explainable artificial intelligence approach for financial distress prediction. Inf. Process. Manag. 59 (4), 102988.

Mark Anthony CAMILLERI, pH.D. (Edinburgh) is an Associate Professor in the Department of Corporate Communication at the University of Malta. He was a Fulbrighter at Northwestern University in Evanston, U.S.A (in 2022). Prof. Camilleri was featured among the world's top 2% scientists in Elsevier's "Updated science-wide author databases of standardized citation indicators" (in the past four years). In 2023, he achieved a global rank (ns) of 3854, and was listed 124th among business & management researchers. He serves as a scientific expert and reviewer for various European research councils. He was recognized for his outstanding reviews by Publons and by Emerald (as he received a Literati award in 2022 and 2023). He is an Associate Editor of Business Strategy and the Environment; Sustainable Development and of International Journal of Hospitality Management, among others.