World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

EXPLAINABLE AI FOR HIGH-STAKES DECISION-MAKING

Sujon Sarkar [1,*] and Lawrencia Yaa Kakra Gaisie [2]

[1] Department of Artificial Intelligence, Touro University, New York, USA.
[2] University of Tulsa, Collins College of Business, 215 Helmerich Hall, 3102 E. Fifth St. Tulsa, OK 74104, USA.
* Corresponding Author

ORCID Details
Sujon Sarkar: https://orcid.org/0009-0008-3319-3204
Lawrencia Yaa Kakra Gaisie: https://orcid.org/0009-0001-1698-3561

World Journal of Advanced Engineering Technology and Sciences, 2026, 21(02), 074–089

Publication history: Received on 06 September 2026; revised on 02 November 2026; accepted on 04 November 2026

Article DOI: https://doi.org/10.30574/wjaets.2026.21.2.0491

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

As artificial intelligence (AI) becomes more prevalent in high-stakes decision-making, it presents a great opportunity to enhance efficiency, accuracy and consistency in critical areas like healthcare, finance, criminal justice, education, and public administration. Many AI models, however, are complicated and opaque, which can create concerns about the transparency, accountability, fairness, and public trust of such systems, especially when they have significant implications for individuals and society. This research aims to investigate how Explainable Artificial Intelligence (XAI) can help to overcome these challenges by making decisions made by AI more understandable, interpretable and accountable to human decision makers and stakeholders affected by AI decisions. It discusses essential methods of XAI such as feature attribution, model-agnostic explanations, interpretable machine learning models and counterfactual explanation, and evaluates their performance in communicating the rationale behind AI generated results. The paper also explores the connection between explainability, accuracy of decision making, algorithmic fairness, supervision to humans and regulatory compliance. Important issues of balancing model complexity with interpretability, ensuring explanation fidelity, reducing bias, and tailoring explanations to the needs of different constituencies are highlighted. This research also takes into account the shortcomings of current explainability procedures, such as the danger of false explanations or the discrepancy between what users believe the model does and what it actually does. The study explores these questions and seeks to provide a framework for assessing explainable AI systems within the context of high-stakes scenarios, focusing on meaningful transparency, contextual relevance, and human-centric accountability. In conclusion, the study stresses the need to move beyond technical considerations and view explainability as a core aspect of responsible AI governance, enabling informed human decision making, trusted AI systems, and fairer decision making.

Keywords: Explainable Artificial Intelligence (XAI), High-Stakes Decision-Making, Artificial Intelligence Transparency, Model Interpretability, Algorithmic Accountability, Algorithmic Fairness, Human-AI Interaction, Responsible AI, AI Governance, Trustworthy Artificial Intelligence.

1 INTRODUCTION

Artificial Intelligence (AI) has its relevance attached to the current society as it plays a significant role in decision making of the people in the organizations and government. AI technologies like machine learning, deep learning, and natural language processing have opened up avenues to analyze vast amounts of data, uncover complex patterns, automate repetitive tasks, and inform decisions that would traditionally take a lot of time and effort. Today, AI is being used in various fields like healthcare, finance, education, criminal justice, insurance, jobs, and government to name just a few. These include activities including disease diagnosis, credit assessment, fraud detection, student performance assessment, recruitment, risk prediction, and allocation of public resources. As AI becomes increasingly prevalent, it has the potential to enhance efficiency, consistency, and decision-making quality. The use of such technologies, however, does raise some questions about the production of their decisions, the trustworthiness of their outputs, and the assessment of their consequences.

AI's role becomes even more crucial when it comes to decisions that greatly impact people's lives, rights, opportunities, safety, or financial health. These decisions are often referred to as high-stakes decisions, as they can result in significant consequences for individuals or communities if made incorrectly or if there is an unfair result. In healthcare, for instance, an AI tool can help health professionals detect patients who need immediate intervention. A credit-scoring model can determine if someone qualifies for a credit, and an automated recruitment system could determine if someone is employed or not. Likewise, predictive systems from fields such as criminal justice or public administration might affect the outcomes of risk assessment, access to services, or distribution of government support. In such cases, the speed or predictive accuracy of an AI system is not enough to assess its performance. One also has to ask if the system is effective and accurate, if the decisions made by the system are fair, if affected individuals can understand the basis for the system's decisions, and if a responsible person or institution can be held accountable if something goes wrong.

The challenge of understanding how some models get to their predictions is one of the major challenges associated with modern AI systems. Many traditional machine learning methods, like decision trees and linear regression, can give relatively simple explanations of the way they make their decisions, if they are designed appropriately. However, there is a class of advanced models, such as complex neural networks and ensemble learning models, that might have a large number of parameters, interdependencies, and nonlinear relationships that are not readily interpretable by humans. While these models may have high predictive accuracy, they may not be readily understandable to the people who use the products of the model. The black-box problem of artificial intelligence is often referred to as this problem. When a system generates a prediction or recommendation, and the rationale for this prediction or recommendation is not easily understandable by the users, the developers, the auditor or any affected individuals.

When AI systems are applied in actions that have a significant impact on people, the black-box problem raises serious concerns. When a prediction is made by a medical professional that a patient has a high risk of suffering from a certain condition, the health care professional should have enough information to decide if the prediction is medically reasonable and pertinent to the patient's situation. If a financial institution refuses a loan due to an AI algorithm, they may be required to provide a reason for their decision and answer whether the applicant got a fair deal. Similarly, someone rejected by an automated recruitment process should be able to grasp why. If there is no explanation, it is harder to: challenge false predictions, uncover inappropriate patterns, identify discrimination or re-think decisions. Lack of transparency can also lead to a lack of trust in AI-driven services, especially if those individuals do not have the means to contest the decisions made regarding them.

In response to these concerns, Explainable AI (XAI) has become an important field of research. XAI is an umbrella term for the techniques that can be used to increase the transparency of AI systems' behaviour, predictions, or recommendations to humans. XAI does not aim to explain the result of an AI but rather to give information that enables users to learn why a specific prediction was generated, what features were the key to that prediction, how the system acts in different circumstances or what changes might result in a different prediction. The explainability is not just a way to show technical facts of an algorithm. It is to give explanations which are suitable to the needs, knowledge, responsibilities and decision making context of those who receive them.

There are a variety of ways to achieve this goal with different XAI techniques. Some methods give explanations for individual predictions and specify what features led to a certain outcome. Local Interpretable Model-Agnostic Explanations (LIME): LIME approximates the behavior of an underlying complex model around a specific prediction, allowing users to gain an understanding of the underlying model. SHapley Additive exPlanations (SHAP) can be thought of as a way to learn about the contribution of different features to a model's prediction, by leveraging concepts of cooperative game theory. Other strategies use counterfactual explanations that portray how a prediction would be different under different conditions. In other cases, a counterfactual explanation for the loan rejection could be that the model would have differed had it considered a different debt-to-income ratio or repayment history if that was a relevant

and realistic alternative. Besides, there are inherently interpretable models such as a carefully-designed decision tree and rule-based systems, which seek to provide understanding of the decision process without using an entirely different explanation method.

These methods are valuable tools to enhance transparency but they are not effective in all cases. An explanation might indicate influential features, but not demonstrate if the model is correct or incorrect. It might make a decision very easy that it overlooks crucial information, or it might give information that is technically correct but not easily understood by the target user. An explanation can also give the false impression of certainty if users think that a good explanation of a prediction means the prediction is reliable. For instance, if it is shown that age, income, or medical history, or any combination of these factors, influenced a decision, this is not sufficient to constitute showing that the model did so well. This therefore needs to be taken into account alongside other fundamental attributes of AI systems, such as the accuracy, reliability, robustness, fairness, privacy and appropriateness for the intended purpose.

The demand for explainability is also closely tied to the increased focus on responsible AI development and governance. It is important for organizations to understand the design, testing, monitoring and implementation of AI systems in the real world. Developers must know the behaviour of their models and organizations must have procedures to investigate errors, document system limitations and manage potential risks. Explanations should facilitate decision makers' ability to make informed judgments, not to blind them to the need for alternatives. Meanwhile, regulators and policymakers must be able to gauge whether the use of AI is compliant with the legal requirements, ethics and responsibilities applicable to those applications. Among the characteristics that are important in an AI-based system are validity, reliability, accountability, transparency, explainability, interpretability, privacy, and harmful bias management, as covered in the National Institute of Standards and Technology's Artificial Intelligence Risk Management Framework. These principles will show that trustworthy AI goes beyond just functionality.

Especially in high stakes situations, human intervention is crucial. AI systems cannot be a replacement for professional judgment simply for their speed in processing data or for making consistent predictions. Their outputs should be assessed in the context of the professional knowledge, relevant evidence, institutional procedures and the situation of impacted people, however. Explainability can aid in this process by facilitating users to challenge recommendations, detect potential issues, and comprehend the constraints of automated systems. But the human oversight needs to be substantial. The fact that users do not have the training or authority to challenge an AI recommendation doesn't necessarily mean that an explanation will prevent harm. Likewise, users can become too reliant on automated suggestions when they think that the AI system might be more objective or consistent than human decision-making. Explainability, therefore, needs to facilitate critical thinking and the right action, not uncritical acceptance of outputs produced by AI.

Also, the ways in which explanations will be given may vary depending on the needs of different parties. Detailed technical information of feature importance, model sensitivity or prediction errors is needed by a software developer who is studying the behaviour of a model. An explanation may be required by a medical professional and should include a connection between an AI recommendation and clinically relevant evidence. An individual impacted by an automated decision may require information about the main reasons for the decision and how to request a review. Documentation to support independent auditing and compliance assessment may be required by a regulator. What works for one group of readers will not necessarily work for another. This places design of human-centred explanations as a key element of the XAI research and practice.

Moreover, explainability is context dependent in terms of how an AI system is applied. A method that gives a good explanation to a fairly straightforward financial forecast may not work for a complicated medical prognosis or a risk assessment for a public sector project. Which explanation methods are appropriate are determined by the consequences of an incorrect explanation, the sensitivity of the data, the expertise of the user, and the opportunity for challenging the resulting decision. For this reason, it is essential to test XAI techniques with benchmarks that take into account technical characteristics as well as their utility for decision makers.

In the context of this background, the current research explores the topic of Explainable Artificial Intelligence for High Stakes Decision Making. It discusses the key methods to represent AI predictions, their pros and cons, and the potential for representing them for the purpose of fairness, transparency, accountability, and responsible human oversight. The study also acknowledges that explainability is not a panacea for all issues related to AI. Rather, it is a key element of a larger strategy for responsible and trustworthy use of AI. The paper aims to explore the technical and human aspects of explainability in order to get a better understanding of how to assess and apply AI systems more responsibly in scenarios where their decisions can have significant impacts.

1.1 Statement of the Problem

A new challenge has emerged with the use of artificial intelligence in high-stakes decisions for people and institutions relying on trustworthy, fair and accountable decisions. While AI systems can analyze complex information and discern patterns that might be hard to spot for a human, they don't always provide explanations that humans can readily understand. This restriction becomes problematic in areas such as healthcare, employment, financial services, education, the criminal justice system, and access to public resources, where AI decisions impact our lives. In these contexts, a faulty or unfair forecast can impact an individual's health, livelihood, rights or future possibilities. Therefore, when making consequential decisions, more is needed than evidence that a model can make accurate predictions. It also demands mechanisms to comprehend, evaluate, challenge and properly control actions that can be taken by that system.

One of the major challenges in this problem is that some of the advanced AI models are black boxes. A complex model could give predictions based on many variables and relationships, which might not be easily understood. If the logic behind the prediction is not described, it can be difficult for users to decide whether the result is backed up by appropriate evidence or whether there are inappropriate patterns in the data. Developers might also struggle to trace the cause of unexpected behavior and organizations might have trouble when dealing with complaints or explaining the rationale behind automated decisions. This opacity can undermine the trust in AI systems, especially in high-stakes scenarios where the reliability of the results is crucial for decision-making.

Another significant issue is algorithmic bias. The patterns that AI systems learn from the data used to create them can be patterns in the data, which may include historical inequalities, incomplete information, unrepresentative samples, or patterns of existing discrimination. A model can replicate or magnify inequitable effects if these inequities are not properly identified and dealt with. For instance, an automated recruitment system that uses historical data on job disparities can be problematic for some groups. A financial assessment model can generate systematically different results for groups that are under-represented in the data's financial history. Likewise, a health-care model could be less effective for some groups if the data that the model was developed or validated on do not adequately represent their characteristics. An explanation method might be used to uncover which factors affect a prediction but it is not equivalent to setting forth a judgement of whether or not the model is fair. Nor does it automatically fix any biased training data, objectives that are not relevant or outcomes that are not equal. The challenge is therefore how explainability can play a role in identifying and understanding bias without using it as a replacement to good fairness measurement.

Lacking transparency also poses barriers to accountability. For many high stakes decisions, the responsibility for the decision can fall to multiple parties including developers, data providers, technology vendors, organizational managers, and professionals relying on AI-generated recommendations. If a system has delivered an undesirable result, it might be hard to decide if the issue was due to bad data, wrong model design, wrong deployment, lack of human oversight, or a misinterpretation of the result. Without explanations, if explanations are unreliable, or if the explanations are not documented, it may be more difficult to investigate the decision. This may also impact someone's ability to comprehend the reasons for a choice and if they are entitled to ask for a review. Explainability is thus closely tied to and related to the larger question of who is responsible for what, how things are monitored, and how the consequences of decisions can be challenged.

To tackle some of these challenges, techniques of explainable AI have been developed; however, their practical applicability is still an issue. LIME, SHAP, counterfactual explanations and interpretable machine learning models are a number of different approaches to communicating the behavior of AI systems. But they have different technical presumptions, computing requirements, stability and applicability for specific applications. Some approaches offer local explanations for specific predictions, whereas others try to describe the patterns throughout the whole model. Some of them give numerical feature contributions that might be helpful for technical specialists but hard for non-technical ones to understand. Others give simple rules or hypothetical cases that are more easily conveyed but that may not contain all the salient features of a decision. Thus, having an explanation method does not guarantee the correctness of the explanation, the fact that the explanation is understood, or that the explanation is useful or appropriate for a specific high-stakes application.

A similar one is the evaluation of explanations. Traditional machine learning assessment focuses more on predictive accuracy, precision, recall, and so on. Such metrics are useful, but are insufficient to determine if an explanation improves a person's comprehension of a prediction, or to improve their decision-making process. The behaviour of the underlying model is not necessarily accurately represented by a technically plausible explanation. It is also important to remember that if the data in an explanation are only slightly altered, then the explanation itself will change, and it is unclear what effect this change in data would have on the resulting explanation. Also, an easy-to-understand explanation could still lead to faulty confidence, if it neglects to mention certain limitations. A range of dimensions must be considered when looking at XAI, such as fidelity to the model, consistency, interpretability, usefulness, fairness implications, and the support of informed human judgment. There is not a single neutral way to evaluate the method,

which makes it difficult to compare methods and make a selection of the methods that are most suitable in different contexts, where the stakes are high.

The second key point is the existence of explainability-human behaviour relationship. Explanations are typically given to enable people to evaluate the AI recommendations better. Yet in some instances, explanations will not enhance decision-making. Users might be misled by technical information, get their attention on the wrong features, or take a recommendation for granted due to its apparent credibility. In some situations, a thorough explanation can help build trust but not the reliability of the decision-making. In some, the user might refuse a valid suggestion due to lack of information or because the information is not what the user expects. The examples illustrate that not every explanation technique can be evaluated on whether it yields a visual display, a feature-importance score, or a written explanation. How the explanation influences the reasoning, confidence and decisions of the users of the explanation is also important.

The process is further complicated by high stakes sector differences. There are a number of distinct sets of requirements, responsibilities and risk levels that healthcare professionals, financial institutions, educational institutions, public administrators and criminal justice agencies face. A medical explanation might need to back up clinical judgement and patient safety while a financial explanation may need to explain the reasons why an individual may be ineligible for a service. A public-sector explanation may require explanations to meet the principles of transparency, procedural fairness and the opportunity to challenge a decision. The differences in these make it impossible to assume any single technique of explaining and any single criterion of evaluation is automatically applicable in all applications. A systematic evaluation would help to determine which approaches are appropriate in specific settings and what constraints need to be accounted for prior to implementation.

A key question examined in this study is the challenge of measuring the technical reliability, explainability, understandability, usefulness and the potential to promote fairness and accountability of explanations generated by the existing explainable AI techniques. The advantages of XAI can be more pronounced if the explanation is of high quality, if the AI system is well designed, if the user needs it, and if the governance framework is appropriate for the use of the technology. If these elements are not properly integrated into an organized approach to assessing them, organizations can end up using explanation processes that seem logical but do not reflect the real risks associated with AI-based decisions.

To solve this issue, the study reviews existing XAI techniques to assess their potential benefits and drawbacks in the high-stakes decision-making process. It aims to establish guidelines for evaluating explanations, look at how accountability and fairness can be integrated into the assessment, and suggest an evaluation system that unites technical performance, human comprehension, and responsible management. The goal is to understand that explainability does not necessarily translate to trustworthiness, but rather what conditions enable AI to play a role in more transparent, fair, and accountable decision making.

Aim and Objectives of the Study

The primary objective of this study is to explore the use of XAI techniques in high-stakes decision-making, their strengths and weaknesses, and to define an evaluation framework to measure their effectiveness in achieving transparency, interpretability, fairness, and responsible human oversight. Moreover, the study acknowledges the need to assess explanations within the context of the choices that they are meant to guide and the audience that depends on them. For this reason, it takes into account the technical aspects of explanation methods as well as the implications they have for decision making in consequential environments.

For the purpose, the study aims at discussing the core ideas and methodologies of XAI, and highlighting the role of various techniques in making AI predictions more explainable. This includes examining model-specific and model-agnostic methods, local and global explanations, feature-attribution approaches, counterfactual explanations, and inherently interpretable models. A study of these methods will give a basis for understanding the way explanations are constructed, the information they convey, and their underlying assumptions that might lead to their use in unsuitable ways in the context of high stakes applications.

The study also seeks to evaluate the effectiveness of the selected XAI techniques based on the relevant evaluation metrics. This includes the question of whether explanations are a true representation of the behavior of the model, whether they are relatively stable in the face of "reasonable" changes in the input to the model, and whether they convey information in a form that users can understand. The assessment also takes into account how far explanations can go beyond just rendering model outputs more visible or transparent as to enable informed decision making. The aim of this evaluation is to uncover the advantages and disadvantages of various approaches, and to determine when specific techniques might be more or less suitable.

Another goal is exploring the connection between explainability, algorithmic fairness, transparency and accountability. The study addresses the potential for using explanation methods to identify influential variables, investigate unexpected predictions, and to gain insight into potential sources of unfair outcomes. It also acknowledges that explanations cannot be used for identifying or fixing biases, and underscores the need to use other complementary practices, including data-quality assessment, fairness testing, model validation, documentation, and independent auditing. This objective aims to shed light on the role of XAI in responsible AI use but not to suggest that an explanation is sufficient to ensure a fair or accountable decision.

The study also aims to explore the extent to which human oversight is necessary in high-stakes decisions involving AI. It reflects the importance of the way explanations can help professionals, decision makers in organisations, affected people and oversight bodies to critique recommendations, acknowledge uncertainty and decide when human intervention is needed. This involves reviewing the potential dangers of relying too heavily on automated deliverables and that explanations could, in some cases, lead to more confidence without corresponding gains in decision quality. Therefore, in the study, explainability is looked at with respect to human understanding, professional responsibility and the processes by which consequential decisions can be reviewed or challenged.

Finally, the study seeks to offer a structured evaluation framework to consolidate technical, human-centred, ethical and governance related aspects. The proposed framework is designed to facilitate the systematic comparison of XAI techniques and help determine and assess explanations in various high-stakes scenarios. The framework will not leave it to any individual metric of explainability, but will take into account several aspects, such as the level of fidelity in the explanation, its stability, its understandability, its utility, its implications for fairness, and the ability to account for it and for human oversight. The study aims to advance the vision of explainability as a metric and tool for responsible AI development and deployment by providing a more holistic understanding of how to measure explainability.

Research Questions

The guiding question for this study is: What is the applicability of Explainable Artificial Intelligence techniques in assisting responsible decision making in high stakes settings? The question exists because it is important to be able to understand whether the explanations generated by AI systems contain meaningful information about the predictions made, and whether it contributes to the decision-making process for the user in a way they could be held accountable for. There is a need to look at technical quality of explanations and their utility under circumstances where decisions might make a considerable difference to individuals and institutions.

The first research question is related to the primary methods employed to provide an explanation to the predictions made by AI models and the distinctions between them. It asks how LIME, SHAP, counterfactual explanations, and inherently interpretable models provide explanations, what information they provide, and how their characteristics impact their appropriateness for high-stakes use. There are differences to be understood because there are different explanation methods, and they may be useful for different explanations, for different decisions, and for different purposes.

The second research question is about how to assess the effectiveness of XAI techniques. It looks at what is deemed suitable for evaluating if an explanation is a correct representation of the model behaviour, and whether this explanation is sufficiently stable, understandable by the intended users, and useful in making decisions. This question also tackles the difficulty of comparing explanation techniques when the output of different techniques is in different forms and when technical measures may not fully reflect the practical utility of the technique.

The third question of the research is on the connection between explainability and algorithmic fairness. It poses questions about the potential usefulness of XAI in detecting problematic factors, learning the difference between model behaviour and human behaviour, and facilitating investigations into biases and discriminatory outcomes. It also takes into account the constraints of explanations to determine fairness and the need to supplement explanations with other strategies to measure data quality, model performance, and differential impacts within relevant groups.

The fourth research question is about the drawbacks and dangers of explanations in consequential decisions. It looks at how unreliable explanations, instability, overly simplified explanations, privacy considerations, and varying understandings among users can impact the reliability of XAI. It also reflects the potential for generating false confidence in explanations, and for users trusting the recommendations generated by AI without questioning them. Knowing these constraints is essential to establishing conditions for the responsible use of explanations.

The fifth research question is about the need for human oversight, accountability and institutional governance for effective use of XAI. It questions how explanations can contribute to meaningful human review, when to intervene, and how decision making processes allow for questioning or auditing of AI-supported decisions. This question

acknowledges that the value of explainability is not just a technical matter, but one that requires access to authority, training, responsibilities, and procedures of those who utilize AI systems.

The last research question is related to the idea of combining the results of the evaluation of XAI techniques into a practical evaluation framework. It queries what technical, human-centred, fairness-related and governance principles should go into any such framework in order to select and evaluate explanation methods in high-stakes situations. This is done so that a consistent method can be developed to determine if an explanation is suitable for its intended purpose and if it is useful for responsible decision making.

These research questions guide the examination of the impact of explainable AI in terms that go beyond the presence of explanation. The questions link the technical aspects of an explanation method with the constraints of the context, be it personal or institutional, within which the explanations must be judged as being worthy of confidence.

Significance and Scope of the Study

This study is significant as it illuminates the role of explainability as a crucial element in responsible AI for high-stakes decision-making. Stakeholders need more than systems that are predicting, as AI is becoming increasingly used in significant processes. They must learn how to interpret the forecasts, evaluate their limitations, identify potential risks and when to rely on human judgment. This research aims to explore how the various XAI technologies can facilitate transparency, fairness, accountability and adequate oversight, while identifying their strengths and weaknesses.

Researchers will find the study to be a well-structured discussion about the main approaches that can be used to understand explainable AI and the metrics that can be used to measure them. There are many approaches to research in XAI, and various definitions of the usefulness of the explanation, as well as various evaluation strategies. The study aims at clarifying the relationships between explanation fidelity, its stability, its interpretability, its usefulness and responsible decision-making by combining technical aspects and human-centred requirements. It can also serve as a tool to point out where additional research is warranted, especially in regard to assessing explanations in a consistent manner in different models, users, and application scenarios. The proposed evaluation framework could be used as a concept for further empirical research, comparative research and for creating context-specific evaluation approaches.

The study underscores the need to incorporate explainability into the development and deployment of AI applications, rather than as a standalone feature added after the fact for an AI developer or system designer. A knowledge of the advantages and disadvantages of various explanation methods can guide developers in choosing methods that are suitable for the model and target users. The study also highlights the need for testing fidelity, stability and clarity of explanations and the need for interaction with predictive performance, privacy and fairness. The factors outlined above can help in creating AI systems that can be better understood and have clearer limitations described.

The study is relevant for organisations that do make consequential decisions supported by AI, in establishing appropriate procedures for reviewing and monitoring the AI-supported decisions. A better understanding of information that an explanation can provide, the conclusions that should NOT be drawn from the information, and the extra checks that need to be performed to find if a system is appropriate for the purpose may be useful to organisations. An organization employing AI for recruitment, financial analysis, healthcare decision making, or public service provision, for instance, might have to identify how users would interpret the explanations of the models, how questionable outcomes can be reviewed, and how individuals impacted by the AI can ask for a human review. The study highlights these factors, which can inform organizations when deciding on the adoption and management of AI technologies.

The research also has policy, regulatory, and auditing implications for those stakeholders who are interested in the accountability of AI. Explainability can help to ensure transparency and oversight; it can make parts of model behaviour more available for inspection. But an explanation does not necessarily constitute evidence that a system meets legal requirements and operates fairly. Technical documentation, risk assessment, independent evaluation, monitoring, and complaints and error correction procedures may also be necessary for effective oversight. This study could inform the conversation on the evaluation of AI systems and the type of evidence required for their responsible use, in addition to the broader governance issues explored in the paper.

The improvements of explainability can also indirectly help individuals affected by AI-supported decisions. Explanations that are relevant, accurate and understandable to the decision are effective in helping people to understand the main reasons for a decision and whether further explanation or review may be needed. This is especially relevant in decisions that shape employment, credit, education, healthcare, insurance, and public service access, decisions made by AI systems. However, the study acknowledges that explanations should take the form of making it accessible to those who receive them. Very technical material may not be useful to the untrained reader in providing

transparency, and a very simple explanation may not give the reader enough information on the limitations. The study thus stresses the significance of helpfulness to the intended audiences in making explanations.

The study is focused on the scope in terms of concepts, methodology, evaluation, criteria and governance issues related to Explainable Artificial Intelligence in the context of high-stakes decision making. It examines some of the most popular approaches to XAI, such as the use of feature-attribution methods, like LIME or SHAP, counterfactual explanations, model-specific methods for interpretation and inherently interpretable models. The general principles, potential benefits, limitations, and applicability of these different methods to distinct decision-making requirements are reviewed. The study also investigates the criteria for evaluating explanations with respect to technical fidelity, stability, understandability, usefulness, fairness, and human oversight.

The study examines high-stakes applications in various sectors such as health, finance, jobs, education, criminal justice and public administration. These sectors are added because the decisions made using AI in these environments could have significant impacts on people's lives and the need for transparency and accountability in such decisions may be greater than decisions whose consequences are of relatively small magnitude. The idea of why the suitability of an explanation method depends on its application context is demonstrated by considering multiple sectors to look for the common challenges. The study does not presume that all sectors are the same, but it looks at the general principles that can be used to guide the responsible assessment of explainability in a variety of environments.

The scope further encompasses explainability with respect to other responsible AI principles, such as fairness, accountability, transparency, privacy, reliability, and human oversight. These areas are taken into account in so far as they relate to the evaluation and use of explanations in practice. For instance, the study explores the potential for explanation to be used to investigate possible biases, acknowledging that fairness assessment needs to go beyond the interpretation of single predictions. Likewise, it recognizes that explainability can be used to aid accountability but does not believe it is a sufficient means to be the basis for legal liability or assurance of regulatory adherence.

This research draws upon the available literature on explainable AI (XAI) to create an evaluation framework. The study will not propose any new machine learning algorithms, or original experimental tests or clinical trials of XAI methods in operational decision-making. Neither will the research include an evaluation of XAI techniques from an empirical perspective. It focuses on examining and understanding the XAI literature from the viewpoint of their deployment for responsible decision-making.

In interpreting these recommendations, one should take into account the limitations of the proposed framework. As the domain of explainability is still evolving, new methods, evaluation approaches and governance standards emerge daily. Explanations also vary from model to model and depend on the dataset, the domain and the user. Some attributes of an explanation, such as understandability and usefulness of an explanation to a user, can only be known by means of a real-world evaluation, as they cannot be deduced from a purely theoretical standpoint. Therefore, the proposed framework should be understood as an approach that needs to be refined and validated to meet the demands of each sector before it can be used as a final evaluation instrument.

Notwithstanding the weaknesses, this research will add a great deal to our understanding of the role of explainable AI in high-stakes decision-making. By investigating the combination of technical and contextual factors in the study of AI, the authors will move beyond the prevailing wisdom that designing an AI system that is ‘understandable’ to a user will make it ‘trustworthy.’ The major contribution of the study is to illustrate that, to be effective, explainability needs to provide a context for its interpretation, and the interpretative information needs to be suitable for the user, in addition to requiring the wider ecosystem in which the decisions made by the AI will be judged to be accountable and fair.

2 LITERATURE REVIEW

Explainable Artificial Intelligence (XAI) has become an important area of research because of the increasing use of AI systems in high-stakes decision-making. Today, AI is being applied across diverse sectors like health, finance, education, criminal justice, and public administration, where the impact of decisions can impact individuals' lives significantly. While AI systems can analyze vast quantities of data and make predictions quickly, some complex models are black boxes, making it hard for users to understand how the decisions are being made. This has led to concerns about transparency, fairness, accountability, and reliability of decisions made with the help of AI.

Explainability, interpretability, transparency and accountability are related yet different concepts. Explainability is about giving understandable reasons for a model's prediction, interpretability is about the ease with which one can understand the behaviour of a model. Transparency is sharing information about the AI system that is relevant, while accountability involves making sure the people developing it, deploying it, and using it have the responsibility to do

that. Miller (2019) points out that explanations from a human perspective are vital, meaning that a good explanation should be relevant to the needs and knowledge of the intended audience.

There are a few methods for enhancing the comprehension of AI systems. The predictions made by SHAP can be decomposed into the contribution of individual features, and LIME can approximate a complex model locally using a simpler one. By showing how changing some of the selected conditions might result in a different prediction, the counterfactual explanation helps readers see how the model changes with those conditions. Surrogate models are models that are used to approximate more complex models, but inherently interpretable models like decision trees and rule learning models are used to make the decision making process easier to understand. Lundberg and Lee (2017) developed SHAP and Ribeiro et al. (2016) introduced LIME. These methods are useful in providing information, but their explanations are not necessarily accurate, stable, complicated, or useful.

The emphasis of the literature is also on explainability in high-stakes applications. Explanations can support investigation in healthcare in relation to a prediction of diagnosis and risk assessment. In the financial world, they can explain the elements that are taken into account when assessing credit and identifying fraud. Explainability in criminal justice can be used to provide scrutiny of risk assessments and decision processes. In education, it may be useful for teachers to comprehend student-performance predictions, and in public administration, it may be useful for the communication and review of decisions regarding public services. But explanations do not ensure correctness, fairness, or appropriateness of decisions. Their value depends on the underlying model, the data they are based on and in what context they can be used.

One of the main issues is the interplay between explainability, fairness and human oversight. The use of AI models can perpetuate inequalities if they are based on biased or unrepresentative data. Explanation methods may help determine which features have an impact and explore unexpected results, but they do not necessarily remove bias. Additional steps such as fairness assessment, data-quality checks, model validation and continuous monitoring are also required. There is also a need for human oversight as users can be misled by explanations or overly trust in the automated recommendations. Effective explanations should enable users to think critically and make informed judgments, not simply accept the output of AI.

While significant strides have been made in XAI research, there are still some challenges. Explanations might not be modelled correctly, might be very sensitive to minor changes in inputs, and might be too technical for intended users. Some explanations can also lead to false confidence, giving predictions a higher degree of accuracy than they really have. There is a need for a rigorous evaluation of the interpretability methods, as suggested by Doshi-Velez and Kim (2017), and the need to think about using inherently interpretable models when they can fit in with the required level of performance for important decisions, as proposed by Rudin (2019). These views emphasize the need for evaluating the quality of explanations, and not taking the presence of an explanation for granted as a sign of trustworthy AI.

The assessment of XAI also is notoriously difficult because there is no one measure that reflects all aspects of explanation quality. Technical assessment can be done to evaluate fidelity and stability, and human-centred assessment can be done to evaluate understandability and usefulness for decision-making. Kim et al. (2024) find significant differences in the evaluation of human-centred XAI applications, highlighting the need for more uniform and in-depth evaluation. Moreover, explainability should also be taken into account along with privacy, accountability, institutional governance, and the impact of incorrect decisions.

The literature, therefore, also shows that there is a need for an assessment framework which integrates technical and human-centred criteria. A framework that evaluates explanations for the behavior of the model should include the following criteria: that the explanations are accurate in representing the behavior of the model, that the explanations are sufficiently stable, that intended users can understand the explanations, and that the explanations can help to make fair and responsible decisions. It should also acknowledge that the explanations depend on the application, and that explanations are still needed by human beings and that human oversight and institutional accountability are still very much required.

To address these challenges, this research investigates key XAI approaches and their applicability to high-stakes decision making. It emphasizes their strengths and weaknesses, and how effective they are for promoting transparency, fairness and responsible oversight. The objective of the study is to create a structured assessment system that takes into account technical reliability, human understanding, practical utility, and responsible governance of AI.

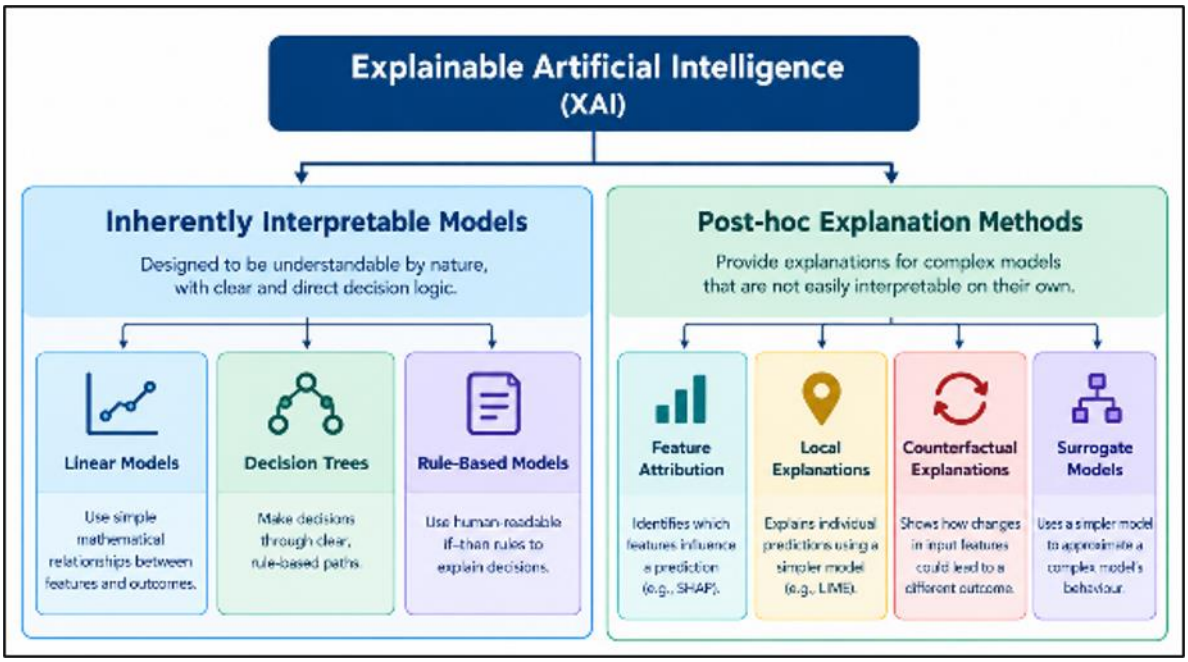


Figure 1: Taxonomy of Explainable Artificial Intelligence Techniques

3 RESEARCH METHODOLOGY

Using a combination of structured literature reviews and comparative analysis, the focus of this study is Explainable AI (XAI) within high-stakes decision environments. Current research is analyzed for common XAI techniques as well as how XAI is utilized in different fields of application. Comparative analysis will identify both similarities and differences between the various XAI approaches for application within decision environments where there is a requirement for both accountability and transparency.

Publications related to Explainable Artificial Intelligence and the intersection of AI with the legal, educational, healthcare, and financial sectors are drawn from different research and scholarly sources, including Google Scholar, ScienceDirect, IEEE Xplore, and the Springer and ACM Digital Libraries. These sources are utilized to locate publications addressing the topic using the search phrases "Explainable Artificial Intelligence," "Interpretable Machine Learning," "High Stakes Decision Making," and "AI Fairness." The literature will also include peer-reviewed journal articles, conference papers, and scholarly reviews addressing the research question.

Studies are selected that present an understandable framework of Explainable AI techniques, as well as those that analyze their performances, and ethical or practical consequences. Studies that don't help to answer the research question or lack sufficient methodological detail will not be used. There will be a focus on both foundational and recent scholarly literature that is pertinent to the topic and of good quality.

The synthesis was organized by first categorizing the literature to form thematic categories like: fairness, evaluation hurdles, human control, real world uses, and interpretability methods. This analysis was complemented by a side-by-side comparison of approaches to XAI that fall within interpretability categories of either post-hoc methods like LIME, SHAP, counterfactuals and surrogate models, or methods that are intrinsically interpretable. For every technique, we compared the methods for use in safety critical settings. Interpretability and trustworthiness of XAI methods were also assessed. Using the identified research gaps, a framework to assess the design and deployment of XAI was proposed. As the entire paper is literature based, its validity and contributions are contingent upon the existing publications that have been reviewed.

4 ANALYSIS AND DISCUSSION

In the first section, it explains the various approaches to explainable AI and discusses the pros and cons of each in terms of its application to high-consequence decision-making. Then they explore how explainable models affect predictive performance, the role of equity and responsibility, and how to construct a robust system for evaluating XAI.

4.1 Comparative Analysis of XAI Techniques

AI systems vary significantly in how they interpret input. Methods can explain individual predictions or provide global interpretations by simplifying the entire model. Techniques to understand an AI model vary depending on the level of complexity, the application and the type of prediction.

SHAP models predictions as a contribution of features. For professionals who are not experts in the domain, it provides a valuable way to interpret decisions in sectors like finance or healthcare. In this approach, predictions that involve a high degree of complexity will require substantial computational resources to generate SHAP values.

LIME (Local Interpretable Model-agnostic Explanations) allows users to interpret the reason behind a complex prediction by fitting a simpler local surrogate model around the prediction. The generated explanation varies based on the data and settings used to construct the local surrogate model. Consequently, LIME generated interpretations are less reliable.

Inherently interpretable models can directly visualize the decision-making process. This provides a clear audit trail and can be a big plus in high-stakes environments, although simpler models lose some detail. These models are particularly good for smaller datasets or for situations where we want to balance model performance with the level of interpretability.

Counterfactuals provide information on how changes in input features affect an outcome. A loan applicant can learn how changes in their input features (e.g. income or credit history) would change the prediction of a loan application. Changes must be realistic and consistent with rules and regulations. Alternatively, a simpler model (or surrogate) can be used to approximate the AI model. Though the model might be simpler, the approximation can misinterpret the original model's logic.

Table 1: Comparison of Explainable AI Techniques

Technique	Strength	Limitation
SHAP	Estimates individual feature contributions to predictions.	Can be computationally demanding.
LIME	Explains individual predictions using a simpler local model.	Explanations may be unstable.
Interpretable models	Provide directly inspectable decision logic.	May not capture every complex pattern.
Counterfactual explanations	Show how changes in input factors could alter an outcome.	Suggested changes may be infeasible.
Surrogate models	Simplify complex model behaviour.	May not faithfully represent the original model.

One of the key things the comparison demonstrates is that a single method cannot be used in all situations. For example, while SHAP and LIME can provide an explanation for more complex models, direct interpretability can be built in as part of the model. When a person needs information that can be easily turned into action, counterfactuals should be used, but when it's about presenting a model to represent a system of greater complexity, surrogate models should be the choice. A choice in methods is only as good as the understanding of the trade-off of the method's technical performance versus what is most useful to those people who will ultimately be impacted by the decision.

4.2 Explainability and Predictive Performance

Both explainability and accuracy are important when designing AI for systems that will impact important decisions. Models that are more complex can produce better results, but these models are much more difficult to interpret. Less complex models are usually more interpretable, but they tend to produce worse results for complex data.

Accuracy alone is not sufficient to evaluate whether a system is ready for use in important decision systems. A system might be accurate even when based on biased data or based on inappropriate data. Similarly, an explanation can be very persuasive and at the same time fail to capture the underlying model.

It is therefore necessary to analyze the tradeoffs between explainability and accuracy. It is recommended that system builders consider the predictive accuracy of a system, but also take into consideration the system's interpretability, the

fidelity of the system's explanations, as well as the practicality of the explanations. In situations where an interpretable model is able to perform adequately, the less complex model is probably preferable. There are benefits to the complex models; therefore, the solution might lie in the system's ability to explain itself more effectively.

4.3 Fairness, Trust, and Accountability

When artificial intelligence affects important areas such as a person's health, education, finances, employment, and legal situations, it must be accountable, trustworthy, and fair. There are various ways that an AI can be transparent about how it reaches a decision, for example, by providing an explanation of the factors leading to a decision, however, this doesn't guarantee fairness.

Let's look at how a lending algorithm can provide a reason for rejecting a loan application and use unfair historical data to determine loan eligibility for certain groups. An explanation can provide the basis for looking into whether a decision was unfair, however, more steps would need to be taken to ensure that fairness has been achieved.

The importance of the understanding and the accuracy of the explanation to its user, cannot be underestimated. While an AI's technical users may need a high level of granularity and may be interested in the contribution of certain features of an algorithm to its decision, a person who has been adversely affected by a decision may only need to know which features were taken into consideration in reaching that decision. An explanation should be relevant and easy to understand to be of any value. Explanations shouldn't be too detailed and complicated. That may leave users feeling even more confused than before they asked for one.

Human review is an essential part of the design and deployment of any system that uses AI. This means the decision makers using these systems should be able to see the recommendations of the AI, review the reasoning behind the recommendation and be able to intervene in cases of AI errors. Organizations also need to define accountability for the decisions made using these systems to ensure the principle of explainability is grounded in something substantive rather than an empty promise of transparency.

4.4 Proposed XAI Evaluation Framework

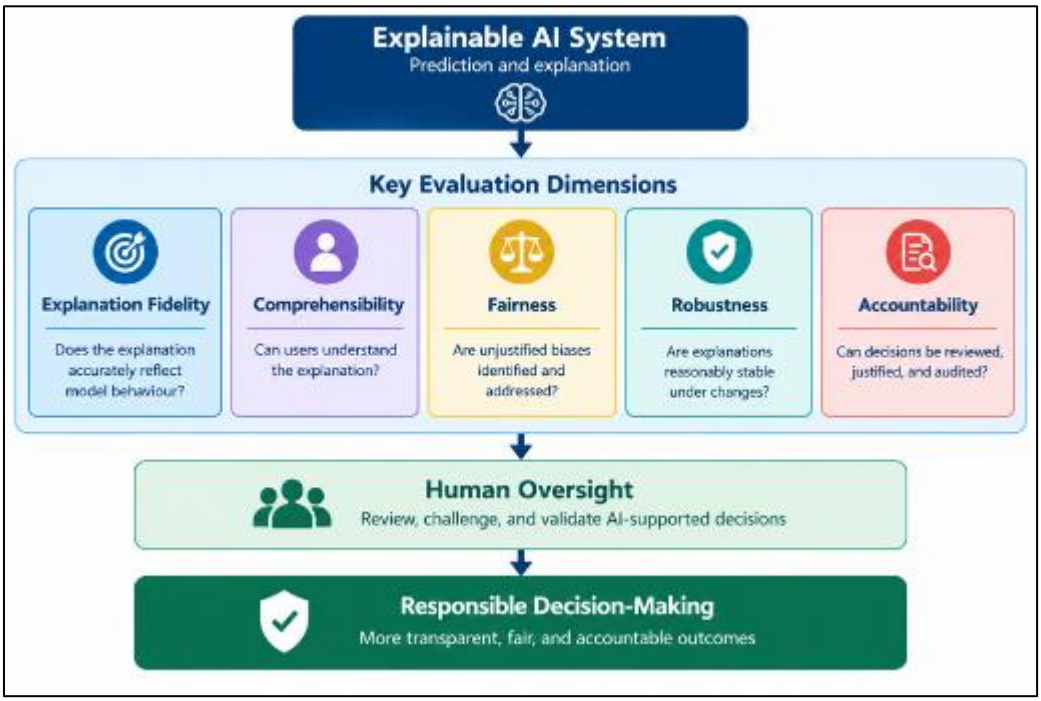


Figure 2: Framework for Evaluating XAI in High-Stakes Decision-Making

This study proposes a framework that examines XAI systems for high-stakes decision-making. Five dimensions were chosen for this framework: explanation fidelity, comprehensibility, fairness, robustness, and accountability. All of these dimensions serve to help people maintain oversight in decision-making and result in more responsible decisions.

The first dimension is explanation fidelity. This refers to the extent that an explanation matches the way the AI model actually behaves. Next is comprehensibility. This dimension considers whether the target users will actually understand the explanation. Fairness considers if the system creates disadvantages that are not justified for certain individuals or

groups. The next dimension is robustness. This dimension examines how much an explanation can be trusted if there are reasonable changes made to the data or model. Finally, the accountability dimension considers if decisions are able to be examined, explained, and traced to the individual or group responsible.

These dimensions are all related. A clear, technically correct explanation can still be unfair. Each of these dimensions needs to be taken into account for an overall assessment of the suitability of an XAI system.

The framework suggests that there are many factors to take into account when making decisions. In order to provide a decision with an explanation that will hold some weight, an explanation needs to be provided that is accurate, clear, fair, and stable. Along with these factors, decisions need to be auditable. In the final stage of decision making, human interaction bridges all of the other stages by being the interpretative step of the decision, and allowing for the opportunity for individuals to dispute the decision, and be more educated in their decisions. The framework needs more work to determine how it would be effective across different areas of the sector, and to prove its utility with more empirical evidence.

4.5 Practical and Regulatory Implications

Using XAI practically in an organization will require explanation methods to be used in every phase of the lifecycle of an AI system (Development, Deployment, Monitoring). Model trainers should write down model limitations, and test for the accuracy of their explanation models against model behavior. It will also be necessary to develop explanations that serve the affected individuals and the decision makers.

Periodic Audits should also be built into the system to handle unexpected behaviors or changing situations like shifting data and shifts in the model's use. There should be protocols to deal with the inevitable problems by reevaluating, correcting, and modifying, or halting the system as appropriate.

There are also many legal constraints on using XAI. Rather than relying on simple explanations, companies must take the time to look at the legal constraints that apply to them. Good XAI will require more than technical tools; it will require ongoing evaluation and oversight.

5 CONCLUSION

The focus of this part is to report the results of the study and the conclusions drawn from it. It contains proposals on how best to employ the principles of responsible XAI. Also included in this part is an analysis of the scope of the study as well as a list of areas in which future research can be conducted.

5.1 SUMMARY OF KEY FINDINGS

In this study, Explainable Artificial Intelligence in high-stakes decision-making is covered including the benefits, challenges, techniques, and the implementation of Explainable AI. Explainability in Artificial Intelligence systems is shown through research and review to be very important, particularly in high-risk sectors such as healthcare, finance, education, and criminal justice, as well as in the public sector.

Explainability is achieved through the use of different techniques in Explainable AI. One example of this is the SHAP technique, where the importance of an individual input in relation to the model's output is shown. A limitation to SHAP, and an example of where a different technique may be preferred, is in the model's overall interpretability, whereas with the LIME technique, a decision is explained through a linear model that represents the input in proximity to the output. Decision trees and rule-based systems are examples of interpretable systems as a whole. One can think about "what if" scenarios in counterfactual systems by visualizing changes to an input that could change an output. There are many different methods and systems available, so it is necessary to consider the context of a decision and who will be using the AI to choose which technique best suits their needs.

Additionally, even with an interpretable system, decision fairness is not ensured, nor is decision reliability. The output explanation could be difficult to comprehend, could be inconsistent and unpredictable, or could simply be an inaccurate representation of the model's process. AI models will retain the training bias. In cases of output explanations, there needs to be an external review that addresses both fairness and validation. Continuous oversight and evaluation is needed, as is system explanation and validity assessment.

RECOMMENDATIONS

Recommendations to aid responsible XAI use in high-stakes environments are made based on the study.

- The first recommendation is for organizations to employ explanation methods that meet the context of a decision and for whom the decision is intended. For example, a clinical judgement in healthcare might require an explanation method that is different from that required by individuals affected by financial or administrative decisions.
- The second recommendation is for AI systems and their explanations to undergo independent validation prior to deployment. The validation should assess performance of the model in making predictions and also in making accurate and stable explanations. Performance in making accurate predictions and explanations should be balanced against the model's complexity. The recommendation is for organizations to use a less complex, inherently interpretable model when it makes accurate predictions and explanations, as compared to using a complex model.
- The third recommendation is for fairness assessments to be done on the AI system throughout the system's lifecycle. The recommendations focus on an assessment of training data, identifying potential sources of bias, and an assessment of whether the system makes outcomes that provide unjustified disadvantages to some groups. The recommendations for making fairness assessments should not be confused with the recommendations for making explanations for outcomes.
- Finally, there should be clear guidelines and protocols in place for human review and the ability for human interventions of systems that provide recommendations in high stakes environments. Finally, it should be clear which systems are in production, what their expected performance and behavioral expectations are, and a regular cadence of review and documentation should occur of the systems' performance.

LIMITATIONS AND FUTURE RESEARCH

This research utilizes existing literature and comparisons instead of empirical data or analysis of actual AI models. Consequently, the research results rely on the literature, which is limited in depth and breadth, and potentially inconsistent. Further, explainability (XAI) techniques have potential for variations across domains, datasets, and applications, making findings not broadly applicable across high stakes environments.

The evaluation of XAI using real world, controlled settings will provide additional data for researchers on the utility and performance of explanations in healthcare, finance, education and others. Studies would also need to compare the performance of XAI across different domains. This, in conjunction with studies of different audiences, would establish the best approaches in practice.

6 CONCLUSION

Explainable Artificial Intelligence is key to having greater trust in decisions made by AI systems. Some methods, such as SHAP and LIME, along with others, provide methods of examining the results of AI predictions. However, these methods, along with others, should be carefully considered, particularly when employed in high-stakes environments.

Ultimately, although there is a lot of merit to AI being explainable, that alone cannot achieve fairness, accuracy and accountability. A multi-pronged approach that includes proper explanation techniques, independent review and verification, fairness reviews and meaningful oversight will need to be adopted, along with ongoing monitoring. If done properly, this will help an organisation to confidently deploy and utilise AI, safe in the knowledge that they are serving their stakeholders.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

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

[2] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Advances in Neural Information Processing Systems (Vol. 31). Curran Associates.

[3] Adebayo, J., Muelly, M., Abelson, H., & Kim, B. (2020). Post hoc explanations may be ineffective for detecting unknown spurious correlation. In International Conference on Learning Representations.

[4] Angelov, P. P., & Soares, E. A. (2020). Towards explainable deep neural networks (xDNN). Neural Networks, 130, 185–194.

[5] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

[6] Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.

[7] Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In Proceedings of the IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI) (pp. 8–13). International Joint Conferences on Artificial Intelligence.

[8] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), Article 832.

[9] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.

[10] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv.

[11] European Commission High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission.

[12] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93.

[13] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. Science Robotics, 4(37), eaay7120.

[14] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4), Article e1312.

[15] Islam, S. R., Eberle, W., Ghafoor, S. K., & Ahmed, M. (2021). Explainable artificial intelligence approaches: A survey. arXiv.

[16] Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. arXiv.

[17] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–14). Association for Computing Machinery.

[18] Kearns, M., & Roth, A. (2019). The ethical algorithm: The science of socially aware algorithm design. Oxford University Press.

[19] Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. In Advances in Neural Information Processing Systems (Vol. 29). Curran Associates.

[20] Kim, J. K., Maathuis, H. M., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. Frontiers in Artificial Intelligence, 7, Article 1456486.

[21] Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–15). Association for Computing Machinery.

[22] Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

[23] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30). Curran Associates.

[24] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

[25] Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 279–288). Association for Computing Machinery.

[26] Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.).

[27] Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. arXiv.

[28] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Article 237, pp. 1–52). Association for Computing Machinery.

[29] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). Association for Computing Machinery.

[30] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

[31] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215.

[32] Sahoh, B., & Choksuriwong, A. (2023). The role of explainable artificial intelligence in high-stakes decision-making systems: A systematic review. Journal of Ambient Intelligence and Humanized Computing, 14(6), 7827–7843.

[33] Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv.

[34] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). IEEE.

[35] Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 56–67). Association for Computing Machinery.

[36] Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (Article 17, pp. 1–9). Association for Computing Machinery.

[37] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

[38] Zhang, Y., Tino, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(5), 726–742.