

Original Article

Human-Centered Explainable Ai for Improving Trust in Automated Financial Decision-Making Systems

Muthukumaran

Research Scholar, India.

Received: 06-09-2026

Revised: 06-24-2026

Accepted: 07-05-2026

Published: 07-12-2026

ABSTRACT

This article proposes a human-cantered Explainable Artificial Intelligence (XAI) framework to improve trust in AI-driven financial decision-making systems, including credit scoring, loan approvals, fraud detection, and risk assessment. Although AI enhances efficiency and predictive accuracy, its lack of transparency raises concerns about fairness, accountability, and user confidence. The proposed framework integrates interpretable machine learning, explanation quality assessment, fairness evaluation, human oversight, and continuous monitoring. It examines various explanation methods and emphasizes that explanations should be understandable, accurate, fair, and relevant to affected individuals. The study provides a conceptual foundation for responsible AI governance in financial services, with further empirical research required to validate the framework's effectiveness.

KEYWORDS

Explainable Artificial Intelligence, Human-Cantered AI, Financial Decision-Making, Algorithmic Transparency, Trust In AI, Credit Scoring, Algorithmic Fairness, Responsible AI, Counterfactual Explanations, Financial Technology.

1. INTRODUCTION

AI has changed how today's modern financial services industry operates through the use of automated analysis of massive amounts of diverse and quickly evolving data. Banks, loan providers, insurance carriers, payment processors, and financial technology firms are using machine learning algorithms to assess creditworthiness, pinpoint potential fraudulent activities, assess potential financial risk, and make informed decisions on investment strategies. In contrast to the traditional rule-based system, data-driven systems allow users to find non-linear connections, recognize patterns in customer behavior, and adjust their analysis based on dynamic financial market conditions. Therefore, these features provide many benefits to organizations seeking to improve operational efficiencies and make better-informed decisions regarding financial transactions.

In addition, as the complexity of algorithms used to create models increases, so does the problem of explaining why a model predicts what it does. While deep neural networks, ensemble models and other types of models with large capacity have the ability to identify complex relationships among the variables in financial data; the inner workings of these models are often incomprehensible to people. When dealing with decisions that will affect many people, simply showing that a model makes good predictions is not enough to show that the model is proper, fair, or trustworthy.

When decisions about money are made, they can have a huge impact on people and businesses. For example, if you get denied a loan, you might not be able to buy a house, go to school, start your own business, or pay for emergencies. If someone gets flagged as a fraudster in error, they may not be able to use their bank accounts for a short period of time. If someone is classified incorrectly in terms of being risky, this could mean that they will receive less favourable loan terms when applying for financial help. If these types of situations occur because of automated systems that don't provide explanations about how decisions were made and/or don't allow people to dispute decisions, then people affected by these decisions may view them as arbitrary or biased.

The emerging field of XAI is focused on developing methods to explain how models behave. Methods in XAI can use representation, feature attribution, examples, counterfactuals, etc., to make model behaviors easier to understand. For example, Ribeiro et al. (2016) created LIME, a method to approximate the behavior of a complex model near a specific prediction. Lundberg and Lee (2017) created SHAP, a unified framework for attributing predictions to individual input features. Both LIME and SHAP are widely used to explain machine learning predictions; however, their effectiveness varies depending upon the model, the data being used, the objective of the explanation, and the intended audience.

However, technical explainability alone cannot create automatic understanding or justified trust. A technical explanation that has many numerical feature contributions can help a data scientist understand what is happening, but it can confuse a person who wants to get a loan. In addition, if you have a technical explanation that sounds good and is written in plain English, it might seem like

you can rely on it, even though it doesn't correctly represent how the model made its decision. Thus, explain ability should not only be looked at as a technical characteristic of the model; it should also be viewed as a design and governance challenge centered on humans.

Human-centered explainable AI focuses on placing the user's needs, abilities, rights, and responsibility in making decisions, at the core of the development process of AI systems. Instead of viewing explanations as being produced as a result of a prediction, human-centered explainable AI looks at who is receiving the explanation, the types of information they need to receive, how those receiving the explanation may take further action based upon the explanation provided, and what methods can be used to contest a faulty decision. Additionally, human-centered explainable AI identifies when users should rely on AI, when they are excessively relying on AI, and when they are rejecting useful forms of automated assistance.

The core concept behind this paper is that the level of reliance placed upon computer-generated financial decisions can be enhanced by utilizing several approaches, including: technical interpretability, clear communications, and evaluation for fairness, organizational accountability and sufficient levels of human control. While it is important to build confidence in computer generated decisions, trust should not be developed without bounds. Rather, systems need to develop "calibrated trust." Calibrated trust means that users will place reliance on the computer-generated decision if the output is determined to be highly accurate, but will use caution and verify or seek assistance from a person when they believe that additional verification or human involvement is necessary.

The study pursues four objectives:

- To examine the principal technical and human factors influencing trust in automated financial decision-making.
- To compare established explanation techniques and identify their limitations in financial applications.
- To propose a human-centered XAI framework integrating interpretability, fairness, user understanding, and institutional oversight.
- To establish evaluation dimensions and future research propositions for assessing the framework's effectiveness.

The article is focused on the use of Responsible AI in Financial Services, specifically for Credit Scoring, Lending and Transaction Risk Assessment. This paper contributes a literature-based conceptual framework as opposed to reporting on new experimental findings. It provides a basis for subsequent empirical studies and implementation research.

2. LITERATURE REVIEW

2.1. Evolution of Explainable Artificial Intelligence

A transition has occurred within the field of Artificial Intelligence (AI) that is moving beyond assessing machine learning systems based solely upon how well they make predictions. Instead, researchers are now looking at the ways in which these systems operate, what they can and cannot do, as well as the potential consequences associated with them. For example, historically, automated decision-making processes were often constructed using explicit rules, decision trees, and/or linear statistical models. While these types of decision paths can typically be traced back to the initial input parameters, there are times when they will have limited effectiveness due to either extreme non-linearity in the underlying relationship between variables, or the presence of a vast number of interacting variables that must be taken into account. On the other hand, the latest advances in machine learning allow for increased predictive flexibility. However, as the complexity of these new machine learning systems increases, it becomes increasingly difficult to determine how specific inputs affect the final outcome.

In their study published in 2017, Doshi-Velez and Kim advocated for studying interpretability via rigorous evaluation rather than viewing it as a subjective attribute of a model. The authors also emphasized the necessity of defining what constitutes interpretability in each application area, and determining if explanations provided by models assist users in completing related tasks. The perspective developed in this study has special relevance to the field of finance since various stakeholder groups within the industry will require varying types of explanation. Developers will want to know how features interact; compliance officers will need proof that all customers are treated consistently; and customers will require understandable explanations when they receive negative decisions.

Ribeiro et al. (2016) created Local Interpretable Model-Agnostic Explanations (LIME) to address the black-box problem. LIME creates local approximations of a complex model with an interpretable surrogate model. This can be used in financial applications to determine which characteristics affected a credit-risk prediction. An explanation is dependent upon how the neighbourhood was defined, how data were sampled from the neighbourhood, and the quality of the surrogate model. Therefore, an explanation does not represent the entirety of the original model.

Lundberg and Lee (2017) created SHAP. SHAP utilizes cooperative game theory concepts of Shapley-values to determine the contributions made by individual features to a model's prediction. The contributions are given as values assigned to each feature, based upon how each feature would have contributed if all other features had been used in the same manner. This enables analysts to examine why a model generated a specific outcome. These contributions may be useful in credit risk analysis, fraud investigations, and model validation. However, while feature attribution can provide insight into what factors were important in generating a model's prediction, it cannot guarantee that the model itself was fair, that the features used are relevant to making a financial decision, nor can it

ensure that the explanation provided will be comprehensible to users who do not possess technical knowledge.

These advancements show us that technical explain ability is a necessary tool to evaluate automated decisions; however, technical explain ability alone cannot create organizational credibility. Therefore, the quality of explanations should be evaluated based upon how the model behaves, what users need, and what could happen as a result of making the decision.

2.2. Human-Centered Design and Trust in AI

The emphasis of human-centered AI is on creating automated systems that are based on what people are able to do, what they need, and what they have to be responsible for. Therefore, when we talk about trust in this type of environment, we are not just saying that someone likes or dislikes a technology. We are making a decision as to whether or not we believe that technology can be used successfully in specific situations based upon the risks involved in using the technology.

Lee & See (2004) studied trust in automated systems, noting that reliance on an automated system should be based upon how well users perceive the system to function, what process it follows, and what it intends to accomplish. While this study was conducted prior to the recent explosion of generative and explainable AI, its findings are still applicable to automated decision making in finance. Users that over-trust automated systems will rely heavily on them and potentially accept false results without verifying them. Conversely, users that do not trust automated systems will ignore valid recommendations from them. These two types of user behaviours can negatively affect decision making quality.

In their review of human trust in Artificial Intelligence (AI), Glikson and Woolley (2020) identified several factors influencing how people develop trust in AI systems. These included system performance, transparency, and the way in which people interact with AI systems. Their results suggest that trust is built upon the interaction between the technological properties of an AI system and people's perceptions of those properties, and is not simply based on the performance of the AI system itself.

Human centered explain ability therefore requires that we pay close attention to who will receive the explanation. For example, a customer whose loan application was turned down needs to know why it happened and what steps they can take next to challenge the decision. On the other hand, a credit analyst may need to know features contributing to the rejection, the degree of uncertainty associated with the decision, and how different scenarios might lead to a different outcome. If you were to provide the same exact technical details to both individuals, you probably wouldn't get the same results in terms of their understanding.

Unfounded confidence should be avoided by explanations as well. An effective persuasive story can create the appearance of credibility for an untrustworthy system; this is especially true if

the interface does not communicate uncertainty or model limitations. Therefore, effective human-centered design must enable users to better evaluate decisions, including enabling users to recognize situations where automated outputs should be questioned.

2.3. Explain ability, Fairness, and Financial Accountability

Algorithmic fairness is a critical component of developing reliable, trustworthy financial AI. Fairness in algorithms can be achieved by removing biases from the data used to train models. However, historical data sets in the financial industry often contain bias against individuals based upon their race/ethnicity, gender, disability status, etc. As a result, if you remove sensitive features from the model (e.g., race) but continue to use other variables that are highly correlated with the removed feature (e.g., zip code), your model will likely still reflect the bias that was present in the original training data. Examples of these proxy variables include geographic location, employment history, and spending habits.

Barocas and Selbst (2016) demonstrated that when making decisions based on data analysis, automated processes that appear to be objective may result in unequal outcomes. The authors' work shows the problems with assuming that if automated processes do not directly include protected characteristics, then those processes will provide equal treatment to all individuals. To develop effective models for finance, developers need to analyse where the data comes from; whether the chosen target variable is appropriate; what types of performance measures will be used to evaluate the model; and what will happen if there are errors in the predictions for different demographic groups.

Explainable AI (XAI) will help with fairness auditing of models by showing which variables are most important for the model's decision making process and revealing previously unknown relationships. But, simply because a model relies on what seems to be a legitimate set of features does not mean that the model will result in equal error rates or systematic disadvantages for some groups of people. Therefore, fairness should be evaluated using many different types of statistical and contextual measurements.

The National Institute of Standards and Technology's (NIST) AI Risk Management Framework defines validity, reliability, safety, security, accountability, transparency, explain ability, privacy, and fairness as key attributes of a reliable AI (NIST, 2023). NIST's AI Risk Management Framework utilizes a life cycle-based approach to assessing AI risks throughout the stages of designing, developing, deploying, and continually reviewing an AI system. Additionally, the AI Risk Management Framework supports the concept that while transparency is required to create a trustworthy AI system, it is alone insufficient.

Financial accountability also demands that institutions make a distinction between providing an explanation of why a model acts the way it does, and justifying a decision based on applicable law and policy. While an explanation can accurately explain how a model relies on income or repayment

history, it may fail to demonstrate that the decision made was in compliance with relevant legal requirements. Therefore, institutions must link technical explanations to formal decision policies, appropriate validation methods, and systems that will allow them to correct any errors.

2.4. Counterfactual Explanations and Actionable Feedback

The concept of counterfactual explanations is based on the idea of describing what would happen to an automated decision-making system if certain input conditions were altered. Wachter et al. (2017) have created a popular approach to providing counterfactual explanations for automated decision-making, which is designed to allow organizations to share relevant information about how they use automated decision-making systems without having to release all details about the system's inner workings.

A counterfactual explanation in credit scoring may provide the applicant with information regarding how a reduction of his/her current debt level or an improvement in his/her payment history would have impacted his/her risk classification prediction by the model. These types of explanations are beneficial since they tie the results of the model to possible future actions.

In addition, developing counterfactual recommendations will require thoughtful consideration. A recommendation system should never recommend changing a static characteristic (such as gender), imply that a user can quickly change their socio-economic status, or recommend spending money in ways that are inappropriate given their financial situation. Additionally, while there might exist a mathematically valid counterfactual recommendation, it may be practically impossible to implement. Thus, this framework provides both a model-valid counterfactual recommendation and an implemented recommendation which has been evaluated for feasibility, fairness and relevance to the organization.

2.5. Research Gaps

The literature reveals several unresolved issues in applying XAI to high-impact financial decisions. First, technical research commonly assesses explanation methods based on their mathematical properties or model specific performance metrics; however, there is generally less assessment on how customers and financial professionals understand these methods. Second, explanation systems are often evaluated independently of fairness, uncertainty communication and appeal mechanisms. Third, the relationship between explanation quality and calibrated trust remains challenging to determine due to many factors affecting trust such as interface design, prior experience, institutional reputation, and potential consequences of a decision. Fourth, if models, financial markets or customer populations change over time then the reliability of explanations may decrease.

Finally, while many frameworks provide general guidelines for transparency and human oversight, they do not offer specific instructions on how to incorporate those guidelines into a full financial decision-making process. This research will fill those gaps by developing a single

framework where each of the five requirements (technical explanation, human understanding, fairness assessment, uncertainty communication, and institutional review) are seen as being related to each other.

3. RESEARCH METHODOLOGY

3.1. Research Design

This research follows a conceptual and literature based approach. This approach brings together existing knowledge about explainable machine learning, human trust in automated systems, fairness in algorithmic decisions and the risks associated with AI to develop a human centered framework for automating financial decisions. In terms of methodology, this is analytical not experimental: the authors do not claim to have collected their own data via customer surveys, institutional datasets, or controlled user studies.

The analysis of financial applications is focused on automated credit scoring, loan approval and transaction risk assessment. These three applications were chosen since each involves the potential of making significant decisions; each uses several types of data inputs; and each requires different types of explanations to customers, analysts and institutional reviewers.

The first step of the study is to identify what are the primary requirements for building reliable and trustworthy AI systems in finance. The second step is to evaluate different types of explanation methods based upon their ability to provide explanations that humans can understand, the accuracy of those explanations, the amount of computing power required to generate them and how useful they would be in practice. In the third stage, we develop a hierarchical human-centered framework using all the identified requirements. Finally, in the fourth stage we derive a set of evaluative criteria and research questions that could be empirically studied in future research projects.

3.2. Proposed Human-Centered XAI Framework

The suggested framework includes six linked layers: financial data management (governance), prediction modelling (prediction), explanation production (explanation), human-focused display (display), oversight and redress (redress) and ongoing assessment (assessment). All of these layers are designed to work together to form an integrated system instead of working independently.

The Financial Data Governance Layer defines how we will govern our financial data (how it comes from, who has access to it, whether it's good quality, etc.), and how to handle sensitive financial information. Data quality is especially critical as no amount of explanation can correct the inaccuracies in income records, outdated repayment histories, or misallocated transactions. In addition to defining the variables that are allowed to make a particular decision, the system should also establish processes for identifying invalid proxy relationships.

The predictive modelling layer will generate the financial prediction or recommendation. In addition to considering how well a model predicts, factors such as its stability, ease of interpretation, operational constraints, and the potential impact of an error should all be taken into account during model selection. A simpler model might be more desirable if it can produce results that are acceptable enough to allow for thorough analysis. If, on the other hand, the use of a more complex model is warranted, there will then need to be further validation and explanatory steps implemented.

The explanation generation layer provides context specific to a particular application and the decisions made by the system. Examples of potential outputs include SHAP values, locally linear surrogate approximations of the decision-making process, comparative analysis of features, confidence estimates, and counterfactual examples. Fidelity of explanations to the original model and stability of those explanations when subjected to small, irrelevant changes in inputs should both be evaluated. It is not assumed that any single type of explanation will always provide the most useful results for all financial applications.

The user interface (UI) translates technical data into understandable information. It uses clear, concise wording; separates influencing variables from cause-and-effect statements; and provides the opportunity for users to obtain further explanation as needed. An initial customer-facing explanation could identify the major factors involved in making a credit decision. An additional, optional technical view would present the numeric contribution of each factor along with validation data.

The oversight and recourse layer is designed to allow individuals to review contested outcomes. This layer provides a way to link the explanations provided by automated systems to human reviews, corrections of incorrect records, formal escalation procedures, and appeals (if relevant). A human review should be a meaningful review process, not simply a ceremonial one. Reviewers must be authorized to challenge the results generated by automated systems, possess the appropriate knowledge and expertise to do so, and have access to the necessary information.

The last level continually assesses how well we predict, the quality of our explanations, whether they are fair, and what the customer experience will be like. The validity of previously created explanations may change with changes in customer populations, financial conditions, data distributions, and/or model versions. Therefore, monitoring must include all aspects of the decision process, not just the model's average prediction accuracy.

3.3. Evaluation Dimensions

To assess the systematic nature of these evaluations, the paper proposes six assessment categories: (1) Explanation Fidelity - does the assessment explain the model's performance; (2) Comprehensibility - will the user be able to comprehend the provided data; (3) Action ability - does the explanation allow for actionable next steps; (4) Fairness - are the errors/outcomes distributed equitably among groups that are most affected by them; (5) Calibrated Trust - does the user have a proper understanding of which results can be trusted, which results cannot be trusted, and which

results may contain errors; (6) Accountability - is there a way to trace back through decisions made, and are there methods for reviewing the accuracy of those decisions.

Each of these components must be assessed individually prior to combining them into a comprehensive evaluation. If you use a single aggregated score, you may miss critical flaws. For example, a very understandable explanation does not make up for an inaccurate or biased decision-making process.

Future empirical tests may include a controlled comparison of an existing automated decision support system (interface) with our proposed human-centered explainable artificial intelligence (XAI) interface. Our participants will be split into two groups, each using a different interface condition. They will then be asked to assess the decisions made by the interface; find potential errors in the decisions; evaluate the level of uncertainty associated with the decisions; and decide whether they wish to follow the recommendations provided or take action to address them. The primary outcome measures of interest will include accuracy of participants' understanding; decision-making speed; identification of errors; measurement of participant's trust in the interface; and evaluation of participants' ability to generate quality appeals.

3.4. Proposed Research Propositions

The conceptual framework generates the following propositions for empirical testing:

- P1: Explanations tailored to users' knowledge and decision responsibilities will improve comprehension compared with explanations presented in a uniform technical format.
- P2: Combining explanation fidelity with uncertainty communication will support better-calibrated trust than presenting feature-importance information alone.
- P3: Integrating fairness assessment and meaningful human review will improve users' ability to identify and contest potentially problematic financial decisions.
- P4: Actionable counterfactual explanations will improve users' understanding of possible next steps when the suggested changes are feasible, relevant, and accurately represented.
- P5: Continuous monitoring of explanations and decision outcomes will improve the long-term reliability of human-centered financial AI governance.

These propositions are hypotheses, not findings established by the present conceptual analysis. Their validation requires controlled studies, representative participants, and appropriate financial decision datasets.

4. RESULTS AND DISCUSSION

4.1. Synthesis of the Conceptual Analysis

The literature based assessment demonstrates that there is a significant difference between providing an explanation and allowing a consumer to assess their decision. While traditional forms of XAI can provide insights into how input data relate to model outcomes, the utility of such insights will depend upon whether the generated insight is accurate, easy to comprehend, and directly

applicable within the context of making a decision. This distinction is particularly relevant in finance; since the provision of an explanation could determine if a customer accepts a negative result, if an analyst reviews a potential error, or if a compliance group determines a systemic issue.

The provided framework is designed to bridge the gap between model-level explanations and how humans understand them, as well as institutional options for recourse. This approach does not assume that providing more information always leads to more trust in a model; instead, it focuses on presenting the right amount of information to the right person. For example, a customer may want a brief overview of why they were approved or denied for a loan. On the other hand, a model validator will likely require in-depth explanations of individual features, performance metrics, subgroup analyses, and the specific version of the model used.

The role specific design is aligned with the National Institute of Standards and Technology's Artificial Intelligence Risk Management Framework; this framework focuses on how to manage trust throughout an AI system's lifecycle as well as taking into account the context in which the systems will be used (NIST, 2023). This proposed framework builds upon the general risk management approach of the framework to create a process for making financial decisions through the direct link of explanation to reviewing, correcting, and continuing to evaluate.

The results of the analysis also suggest that transparency, fairness, and trust are distinct from each other. Transparency is about having enough information available to people; explain ability is about being able to tell why a system works the way it does; and fairness is about whether the output and the process of making decisions are fair based on applicable ethical, statistical, and legal standards. A system can be transparent in explaining itself while at the same time creating unfair results. Similarly, a fair and correct model can be very difficult for customers to comprehend. To create systems responsibly, we need to focus on each of these areas.

4.2. Comparison of Explain ability Techniques

The various ways to explain the predictions of a machine learning algorithm depend upon what type of analysis is needed. Models that are inherently interpretable (such as properly restricted decision tree algorithms and logistic regression models) have a direct relationship between the model architecture and its predictions. The primary benefit of these types of models is that they allow for the generation of explanations directly from the model itself rather than through an external approximation method. Additionally, in many financial application areas, using models that are inherently interpretable simplifies both auditing and documenting the results produced by the algorithm. Whether or not these types of models will be effective for predicting outcomes depends largely on how complex the underlying relationships being modelled are, as well as how the models were constructed.

The use of Post-Hoc methods like LIME and SHAP can provide explanation capabilities for highly sophisticated models. Specifically, LIME can be used to locally approximate a decision

boundary in a complex model, whereas SHAP provides the user with the ability to determine how each feature contributes to the final outcome of the model. This type of flexibility is especially beneficial when an organization must use a very complex predictive model. However, both of these methods must be interpreted carefully. In particular, a local approximation may fail to accurately describe the complex decision boundary of a model. Additionally, a feature attribution value will vary depending upon what assumptions are made during model creation and how relationships between different input features are handled.

Counterfactual explanations offer a distinct type of data by identifying modifications to a system that may result in an altered outcome. One of the main benefits of counterfactuals is that they allow users to connect the analysis of a system's output (predictions) with possible courses of action. However, a user should understand that a counterfactual is not a guarantee that a financial institution would approve a loan if one specific variable was modified. In addition to being based on eligibility criteria, policies and/or other factors that exist outside of the model, actual decisions made by financial institutions can also be influenced by multiple factors. The following table summarizes the comparative implications of these methods.

Table 1: Comparison of Explain ability Techniques for Financial Decision-Making

Technique	Primary Function	Main Advantage	Principal Limitation	Suitable Financial Application
Logistic regression	Directly models relationships between features and predicted outcomes	Relatively easy to inspect and document	May not capture complex nonlinear relationships without extensions	Credit scoring and risk estimation
Decision trees	Represents decisions through sequential rules	Provides understandable decision paths	Large or deep trees can become difficult to interpret	Eligibility screening and policy support
LIME	Approximates a complex model near an individual prediction	Model-agnostic local explanation	Results depend on sampling and local approximation quality	Individual loan-risk review
SHAP	Attributes prediction differences to input features	Supports local and aggregated feature analysis	Attribution does not establish causality or fairness	Credit-risk analysis and fraud investigation
Counterfactual explanations	Identifies alternative inputs associated with a different prediction	Can support actionable feedback	Suggested changes may be infeasible or misleading	Loan feedback and customer guidance
Example-based explanations	Uses similar cases or representative examples	Can make abstract model behaviour more concrete	Similarity measures may hide important differences	Customer support and analyst review

Note. This comparison synthesizes the methodological characteristics discussed by Ribeiro et al. (2016), Lundberg and Lee (2017), and Watcher et al. (2017). It is a qualitative analytical comparison, not the result of a new benchmark experiment.

As shown by the comparison, when selecting an explanation technique, financial institutions need to consider the type of decision being made, the requirements of the individual receiving the explanation, and what will happen if there is an error in the decision. In some cases, a hybrid approach may be most suitable; i.e., an interpretable model can be used to describe the primary reason for making a decision, a post-hoc method can be used to analyse a decision in greater detail, and a counterfactual with appropriate constraints can be used to explain what action could be taken next. However, any such hybrid approach must be tested for accuracy and consistency.

4.3. Human Understanding and Calibrated Trust

Trust, when it is calibrated, is treated as a major output rather than attempting to maximize user confidence. A system that will encourage users to accept automated recommendations without question can provide operational efficiency in the short-run but can result in greater potential for error that is not identified. A system that provides uncertainty regarding each recommendation and does so in a manner that lacks adequate context can overwhelm users and cause them to ignore valuable information.

The appropriate amount of confidence in the system requires the user to understand what the system does well and what it does poorly. In credit scoring, for instance, a description would clearly differentiate between an estimated level of risk provided by the model and a definite statement regarding whether or not the applicant will pay back their loan. A prediction is merely an estimate of risk based upon the data used to train the model, along with the assumptions made when creating the model; therefore, a prediction cannot provide certainty about what will happen in the future.

The design of the explanation interface may help to define these two aspects of the explanation process. It will allow you to present the limitations of your analysis, the primary factors causing your analysis, and any other opportunities for review. When you include confidence values on your explanation interface, you must clearly explain what they represent. Additionally, you need to calibrate them if they are going to provide useful information. You do not have to call a model score a probability unless you know for sure that the model being used provides an option to calculate a probability.

Trust varies among users. Experienced analysts are more likely to identify model limitations; however, they may also become overconfident due to repeated use. Users who are not as technically literate may see automated decisions as institutional judgments that cannot be challenged. Thus, when designing explanations, it is necessary to consider the technical knowledge level of the user and the user's potential for action based upon the provided information.

4.4. Fairness, Bias Detection, and Institutional Accountability

A major finding of the study is that institutional explanations of automated decisions should be linked to an ongoing assessment of how fairly the automated decision-making process affects different groups. For example, if a credit-risk model consistently produces more errors when applied to members of a specific group than it does for other groups, then showing each member of the affected group what features contributed most to their automated decision will not alone address the underlying cause of the disparity. Instead, the institution using the automated decision-making system must determine if the disparity is due to data-quality issues; poor choices about which features to include in the model; unequal representation of the groups in the training data; poor settings for decision thresholds used in the automated decision-making system; or some other systemic factor.

Fairness should be assessed with a variety of metrics that are relevant to the use of the algorithm. As an illustration, if an algorithm is being used to detect fraudulent activity, then false positive/negative rates would likely be a consideration. On the other hand, if an algorithm is being utilized as a credit risk assessment tool, then approval rate, calibration and error distribution would likely be considerations. There is potential for conflict among the metrics, as there isn't a single metric that will guarantee fairness across all conditions of operation. Therefore, when interpreting the results of each metric, you need to keep in mind the objective of the decision, any relevant law/regulation that applies, and what type of errors could occur.

The framework also applies fairness assessment during the model development phase and the model implementation phase. Prior to implementing a model, institutions must identify appropriate subgroups, investigate possible proxy variables, and document any limitations. After a model is implemented, institutions must track changes in the effectiveness of the model, review complaints, and determine if the model is still applicable to the target audience.

Human oversight provides another layer of protection. When a valid review process is used, people who have the necessary knowledge can review decisions made by computer systems that are in question; they can find and fix errors in data used to make those decisions; and they can determine when computer system outputs do not align with policies established by organizations, or with other factual evidence related to the decision-making process. Additionally, human reviews should produce feedback so that improvements can be made to the computer systems. For example, if there are many complaints about one aspect of the system (such as the way it produces explanations), this may mean there is something wrong with the computer model, the interface for using the system, or the decision-making policies themselves.

Importantly, while human review will never completely eliminate the potential for bias within algorithms. Human reviewers may be too dependent upon automated recommendations, and/or may make the same assumptions as the algorithm. Consequently, oversight of these algorithms must contain: clear decision making power, adequate training, and independent review

when required, and documented evidence of why a reviewer either accepts or rejects an automated decision.

4.5. Proposed System Architecture

Figure 1 illustrates the conceptual architecture of the proposed human-centered XAI framework. The architecture represents a design recommendation developed through literature synthesis, not an implemented production system.

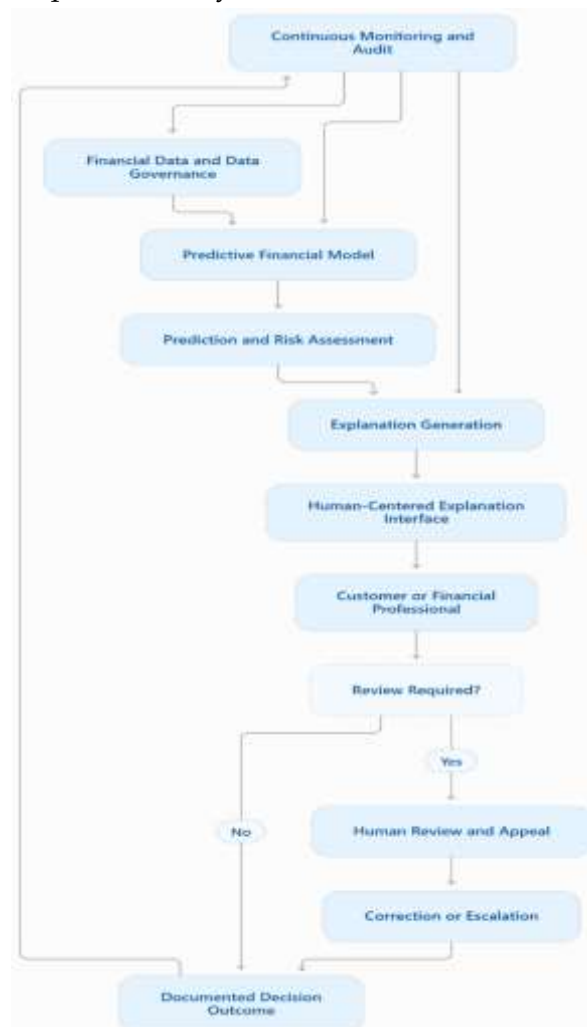


Fig 1 - Proposed Architecture of Human-Centered Explainable AI for Financial Decisions

Source: Author's conceptual framework.

The architecture illustrates that explaining the decision-making process is not the final step of the decision-making process. The results generated by the explanation system will need to be understood by a person who has access to a review system, as well as evaluated continuously using monitoring systems. Auditing and disputes about decisions will help improve how data is governed, the design of models, and the quality of explanations.

The architecture will also be able to provide traceability. Each time an organization makes a significant decision, it should record the version of the model used; the inputs (data) used by the model; what was predicted by the model; how the organization explained the model's output; which policies applied when making the decision; and if someone had to override the model's output. The records provided must comply with the organization's privacy, security, retention, and other regulatory guidelines. These records can then be used by the organization to determine if the model functioned as expected and if the explanation given of the model's output accurately reflected how the system worked.

4.6. Evaluation Framework and Expected Outcomes

To operationalize the conceptual framework, the study proposes a multidimensional evaluation approach. The dimensions in Table 2 can be used to develop a research protocol, an institutional assessment checklist, or a comparative evaluation of alternative XAI interfaces.

Table 2: Proposed Evaluation Framework for Human-Centered Financial XAI

Evaluation Dimension	Illustrative Measure	Evaluation Method	Desired Outcome
Explanation fidelity	Agreement between explanation and model behavior under relevant perturbations	Local fidelity tests and stability analysis	Explanations accurately represent the model
Human comprehension	Correct identification of principal decision factors	User testing and structured comprehension tasks	Users understand the stated reasons
Actionability	Feasibility and relevance of suggested next steps	Expert review and user assessment	Guidance is realistic and useful
Fairness	Relevant subgroup error rates and outcome disparities	Disaggregated performance and fairness audits	Unacceptable disparities are identified and addressed
Calibrated trust	Relationship between system reliability and user reliance	Controlled experiments measuring appropriate acceptance and rejection	Users rely on outputs in proportion to their reliability
Accountability	Completeness of decision records and effectiveness of review	Audit trails, case reviews, and appeal-process assessment	Decisions can be investigated and challenged
Privacy and security	Unauthorized disclosure risks and compliance with data-handling controls	Security testing and privacy review	Explanations do not unnecessarily expose sensitive information
Operational robustness	Performance and explanation stability after data or model changes	Drift monitoring and periodic revalidation	The system remains suitable over time

Note. These are proposed evaluation dimensions and illustrative measures. The framework does not assign numerical scores or claim empirical improvements.

Figure 2 presents the proposed relationship among explanation quality, fairness, human understanding, and institutional accountability. These dimensions are complementary rather than interchangeable. The model is intended to guide evaluation and hypothesis development, not to establish a statistically validated causal relationship.

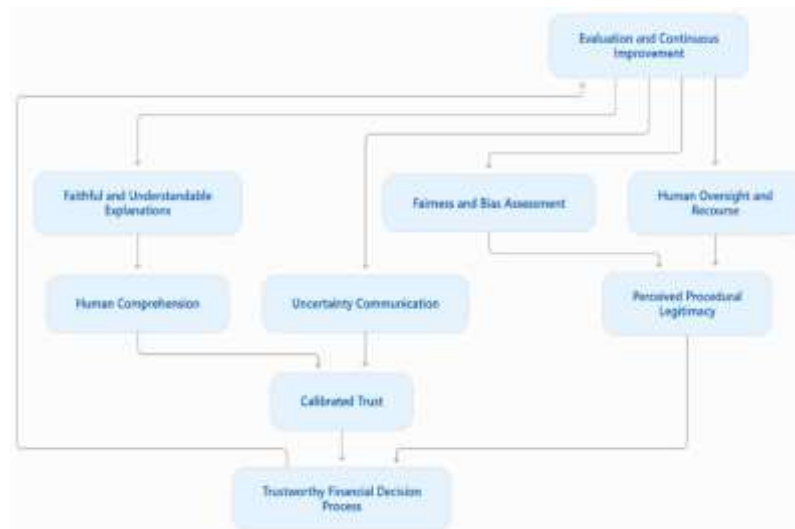


Fig 2 - Conceptual Relationship between Human-Centered XAI and Trustworthy Financial Decisions

Source: Author's conceptual framework.

The above diagram illustrates that trustworthiness is a product of multiple interacting factors. Explanations that are easy to comprehend can assist with understanding, fairness assessments can increase the validity of decision making processes, and communicating about uncertainty can provide an opportunity for users to differentiate between accurate predictions and those that may be inaccurate. Human oversight provides a method to address errors and contest decisions. Finally, continuous improvement connects all four of these components through feedback and assessment.

Expected results for this strategy do not necessarily include higher average trust levels; however, expected benefits include greater awareness regarding how users can rely on automation with confidence, better recognition of errors in automated decision-making and additional avenues for corrective actions. Expected results must be demonstrated via empirical evaluations, and cannot simply be assumed due to the inclusion of explanation features.

4.7. Practical Implications for Financial Institutions

The potential use of this framework could have significant impacts on financial services firms in areas such as technology development, service delivery design, and governance. Those developing models would need to consider the technical requirements for explanation fidelity; in addition to testing if explanations are consistent across various scenarios. Those who develop products would also need to assess whether consumers understand the explanations provided, with no prior understanding of statistical concepts. Those who manage risk and compliance functions would need

to verify that explanations given are consistent with the firm's written policy and regulatory requirements.

Institutions must use explanation interfaces with care, and not as a replacement for developing models in a responsible manner. A system using poor-quality data, or that results in serious disparities among individuals, cannot be rendered credible merely by providing its predictions in clear language. Furthermore, even if an explanation is technically accurate, it cannot make up for a lack of accessibility to an appeals process, nor can it make up for an institution's failure to correct inaccurate customer information.

In cases where decisions are made for systems that interact with customers, explanations of those decisions must be short, clear and practical. When an explanation cannot be completed within one brief sentence, the institution making the decision may provide deeper levels of detail. An initial explanation should include the primary reason(s) behind the decision. As a user requires more information, the institution may provide further details on how the decision was reached and what alternative actions they can pursue if they disagree with the decision. A multi-layered approach to explaining decisions allows institutions to avoid overloading users with too much information, yet still allows users to gain access to more detailed evidence supporting the decision.

For the organization's internally-focused financial professionals, their explanation systems should offer them options to investigate how models behave, compare what they predict, identify anything out of the ordinary in their results, and record when an analyst overrides a result. Analysts need to be properly trained on the limitations of feature attributions, counterfactual explanations, and model scoring. Finally, the organization needs to periodically assess whether the use of such tools helps to make better decisions or simply allows analysts to process information faster.

The framework applies to both large financial organizations and smaller fintech providers. However, implementation costs and technical resource availability will likely vary. Organizations that have restricted resource capabilities should start by developing interpretable baseline models, documented decision rules, simple fairness tests, and a clearly defined review process. As model complexity grows and operational needs grow more complex, organizations can implement more sophisticated explanation systems.

4.8. Limitations of the Study

The findings presented in this article represent concepts that require interpretation. A controlled experiment, user survey of financial services, and/or a benchmark evaluation with real-world financial data was not conducted to evaluate the proposed framework. Therefore, the article cannot determine if there is an improvement in trust, fairness, understanding, and/or operational performance.

The study acknowledges that the way individuals make financial choices varies by country/region and application. The criteria used to make credit decisions, underwrite insurance policies, identify fraudulent activity, and provide investment services are different. For example, a method of explaining how a decision was made that works well in one situation might not work as effectively in another, especially if there are multiple stakeholders who want different amounts of detail or if sharing information about a decision creates issues with privacy or security.

Finally, when providing an explanation, this can also create additional risks. If the design of the explanation is poor, it can lead to revealing sensitive information; it can lead to strategic use of fraud detection systems; and it can create a false sense of certainty. This means the framework must have contextual validation, and we cannot assume that if a system uses components from the framework, that the system is automatically trustworthy.

5. CONCLUSION

Explainable artificial intelligence (XAI) is emerging as a potentially powerful mechanism for improving the transparency, usability, and accountability of automated financial decision making. While XAI has been shown to offer many benefits to automated financial decision making, this paper illustrates why explainability cannot simply involve adding a feature importance chart or natural language explanation to a current machine learning model. In order for an automated financial decision-making system to provide an explanation, there must be a strong connection among the technical performance of the model, what information is provided to the end-user, the fairness of the resulting decision, and how that decision may be reviewed or challenged by the organization.

The research presented in the literature review provided the groundwork for integrating interpretable models, post-hoc explanation methods, counterfactual reasoning, and human-centered interface design. LIME and SHAP are two tools that can be used to understand how a model makes its predictions; counterfactual explanations can relate model outcomes to possible variations in input values. However, both of these types of explanations have different assumptions and limitations. Therefore, prior to using them to make significant financial decisions, it is important to evaluate the extent to which their explanations are faithful, stable, understandable, and relevant to the application area.

This paper makes its primary contribution by providing a six-layer conceptual framework relating to data governance, predictive modelling, explanation generation, human-centered presentation, oversight and recourse, and continuous evaluation. The framework states that financial institutions must create explanations tailored to individual audiences and allow for significant opportunity for humans to intervene in the process. Additionally, the framework provides six key evaluation dimensions (fidelity, comprehensibility, action ability, fairness, calibrated trust, and accountability) which may serve as foundation for subsequent empirical studies.

A second implication of this point is that as a result of trust in financial artificial intelligence being built upon justification and calibration of users' reliance on systems rather than an absolute level of user confidence, then the process of providing explanations for decisions can assist users to recognize the relevant aspects of a decision, understand the uncertainties inherent in making a decision, identify potential sources of error within the decision-making process, and assess whether there is a need for additional reviews of the decision. Additionally, transparency must occur with proper data governance, evaluation of fairness, protection of privacy and accountability for the organization utilizing the system.

The model has yet to be empirically tested and will require further evaluation in order to assess how well it can support responsible financial decision-making in a variety of real world applications. For example, testing the effectiveness of the model with diverse samples (e.g., different demographics) and realistic examples of model results could help determine if there are differences in how individuals perceive the model's performance. As long as we do not have empirical data regarding the value of this model to financial decision-makers, we must treat this proposed structure as a starting point for conducting research into the potential of explainable AI to provide better support to consumers when making financial decisions and as a guide for developing models that incorporate fairness and accountability.

6. FUTURE SCOPE

In the future, there are many ways that researchers may build upon this study to evaluate how human-centered XAI will perform in actual financial environments. One way is to conduct controlled user studies to empirically validate the differences between traditional automated decision making interfaces and human-centered explanatory interfaces. Researchers could investigate if human-centered explanations enhance customer comprehension, error detection capabilities, and proper reliance on systems by customers, financial analysts, and compliance professionals. These studies should evaluate both positive and negative aspects of human-centered explanations, such as if they inadvertently cause customers to become overly confident in their decisions.

In addition to the previous point, there exists a large area of research that involves developing adaptive and personalized explanations. A user interface designed to explain complex processes would have the ability to adapt its level of technical detail based upon a user's current knowledge base, their professional role, and what they are attempting to accomplish at a given moment. Although adaptation is beneficial, it must not prevent users from obtaining the information required to make informed decisions; nor can it result in inconsistent explanations provided to users in similar situations. Therefore, further study is necessary to assess the manner in which explanations can be adapted to provide a consistent experience across all users, as well as ensure procedural fairness.

Third, causal and counterfactual explainability should be explored more fully. Systems developed in the future may allow us to differentiate between what we observe (associations) and what can actually be altered by customers. In lending, this is particularly relevant since a relationship

between an individual applicant's attributes and his/her risk score does not necessarily imply that if he/she changed one of those attributes, the lender would then approve his/her loan application. Causal explanations, if based on explicitly tested assumptions, could enhance the value of financial advice.

The fourth area of focus for fairness-aware explanation evaluation is whether or not explanations can assist in identifying disparities in treatment among different demographic and socioeconomic populations. When investigating these issues, it will be necessary to account for the constraints of existing data, privacy considerations, conflicting metrics used to measure fairness, and the legal framework that applies within each jurisdiction. Additionally, the quality of an explanation should be assessed independently of the fairness of the decision being made, as it would be incorrect to assume that a decision which is understandable to the individual impacted by it is inherently fair.

Fifth, ongoing assessment of explanations as part of the overall management of a Financial Artificial Intelligence (AI) system's life cycle needs to be a core element. Changes to the model, changes to how customers behave, changes to the economy, and new ways people commit fraud will all impact whether your predictions remain accurate, and whether your explanations remain reliable. Systems being developed today could monitor the distribution of explanations, determine which decision factors are inconsistent, and provide alerts for reviews once a certain threshold has been reached.

The final area that deserves specific focus in diverse service settings is the provision of multilingual and accessible financial explanations. Financial explanations should accurately explain financial terms to people who speak different languages and have different degrees of literacy, while maintaining the same underlying rationale for making a decision. Research into how multilingual users interact with financial explanations, what is required for accessibility, and how well people understand digital technology would assist in developing more inclusive design standards.

These directions can move human-centered XAI from a conceptual design principle toward an empirically evaluated approach for improving the responsible use of AI in financial services.

REFERENCES

- [1] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671-732. <https://doi.org/10.15779/Z38BG31>
- [2] Obannagari, C. K. R. N., & Nangi, P. R. (2021). Predictive Database Performance Optimization Using Machine Learning-Driven Query Workload Modeling. *International Journal of AI, BigData, Computational and Management Studies*, 2(4), 107-116.
- [3] Mullapudi, P. N. (2025). Pricing optimization across domains: A comparative review. *International Journal of Emerging Research in Engineering and Technology*, 6(3), 16-33.
- [4] Y. Syed, "Enterprise Observability and Software Delivery Performance: From Telemetry to Engineering Intelligence", *IJADSMC*, vol. 5, no. 2, pp. 01-08, Sep. 2022, doi: 10.67228/30713498/IJADSMC-2022PIIKP2M.
- [5] Nangi, P. R. (2022). Multi-Cloud Resource Stability Forecasting Using Temporal Fusion Transformers. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 3(3), 123-135.

- [6] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. <https://arxiv.org/abs/1702.08608>
- [7] Fung, W., Guo, L., Hilliard, D., Casadei, A., Ratan, R., Bhaduri, S., ... & Rakshit, A. (2025). OpComm: A Reinforcement Learning Framework for Adaptive Buffer Control in Warehouse Volume Forecasting. *arXiv preprint arXiv:2512.19738*.
- [8] Nangi, P. R., & Obannagari, C. K. R. N. (2026). Federated ai-driven query optimization for distributed cloud databases. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 7(1), 91-99.
- [9] Syed, Y. (2024). DevSecOps and Enterprise-Scale CI/CD Pipeline Engineering: Governing Delivery Velocity, Embedded Security, and Release Automation Across Polyglot Environments. *International Journal of Scientific Research in Science, Engineering and Technology*, 11(5), 475-484. DOI: <https://doi.org/10.32628/IJSRSET2513156>
- [10] Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627-660. <https://doi.org/10.5465/annals.2018.0057>
- [11] Nangi, P. R., & Obannagari, R. N. CK, & Settipi, S.(2022). Predictive SQL Query Tuning Using Sequence Modeling of Query Plans for Performance Optimization. *International Journal of AI, BigData, Computational and Management Studies*, 3(2), 104-113.
- [12] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392
- [13] Obannagari, C. K. R. N., & Nangi, P. R. (2021). Predictive Autoscaling for Kubernetes Using Multivariate Time-Series Forecasting. *International Journal of Emerging Research in Engineering and Technology*, 2(2), 73-83.
- [14] Sunkara, R. (2023). Cost-Optimized Energy Compliance Testing for Smart TV Streaming Devices: Achieving Milliwatt-Precision Power Measurement at Sub-One-Thousand-Dollar per Setup. *American International Journal of Computer Science and Technology*, 5(6), 54-59. <https://doi.org/10.63282/3117-5481/AIJCST-V5I6P105>
- [15] Nangi, P. R., & Obannagari, C. K. R. N. (2026). LLM-Driven Database Administration and Natural Language Query Interfaces. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 7(1), 410-418.
- [16] Syed, Y. (2025). Generative AI Integration and RAG-Based Enterprise Tooling: Architectures, Techniques, and Organizational Impact. *International Journal of Scientific Research in Science, Engineering and Technology*, 12(6), 600-606. DOI: <https://doi.org/10.32628/IJSRSET2613255>
- [17] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30 (pp. 4765-4774). Curran Associates, Inc. <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions>
- [18] Nangi, P. R., Obannagari, C. K. R. N., & Settipi, S. (2022). Enhanced Serverless Micro-Reactivity Model for High-Velocity Event Streams within Scalable Cloud-Native Architectures. *International Journal of Emerging Research in Engineering and Technology*, 3(3), 127-135.
- [19] Pavan Nithin Mullapudi (2024). An Adaptive Explainability Framework for Machine Learning Predictions of Deals in Cloud Computing . *International Journal of Multidisciplinary Research and Growth Evaluation (IJMRGE)*, 5(3), 1027-1034. DOI: <https://doi.org/10.54660/.IJMRGE.2024.5.3.1027-1034>
- [20] Nangi, P. R., Obannagari, C. K. R. N., & Settipi, S. (2025). AI-Driven Data Engineering Pipelines for High-Velocity Analytics Using Serverless Architectures and Event-Stream Intelligence. *International Journal of AI, BigData, Computational and Management Studies*, 6(3), 74-83.
- [21] National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- [22] Nangi, P. R., Obannagari, C. K. R. N., & Settipi, S. (2022). Self-auditing deep learning pipelines for automated compliance validation with explainability, traceability, and regulatory assurance. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 3(1), 133-142.
- [23] 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. <https://doi.org/10.1145/2939672.2939778>
- [24] Nangi, P. R., Obannagari, C. K. R. N., & Settipi, S. (2025). Quantum-Enhanced Optimization Models for Large-Scale Security Policy Evaluation in Distributed Cloud-Native Systems. *American International Journal of Computer Science and Technology*, 7(6), 46-56.

- [25] Syed, Y. (2023). Polyglot Persistence and Enterprise Data Migration: Architectures, Consistency, and Practice. *American International Journal of Computer Science and Technology*, 5(3), 61-66. <https://doi.org/10.63282/3117-5481/AIJCS-T-V5I3P106>
- [26] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841-887. <https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf>
- [27] Nangi, P. R., & Nala Obannagari, C. K. R. (2026). Quantum-Classical Hybrid Algorithms for Complex Query Planning: A Scalable Computational Framework for High-Dimensional Decision Optimization in Heterogeneous Data Ecosystems. *International Journal of AI, BigData, Computational and Management Studies*, 7(2), 328-336. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V7I2P143>
- [28] Virendra Jangid. (2025). *AI-Powered Cross-Border Securities Processing and ADR Platform Optimization*. *International Journal of Leading Research Publication*, 6(11). <https://doi.org/10.70528/IJLRP.v6.i11.2233>
- [29] Nangi, P. R., & Obannagari, C. K. R. N. (2024). High-Performance Distributed Database Partitioning Using Machine Learning-Driven Workload Forecasting and Query Optimization. *American International Journal of Computer Science and Technology*, 6(2), 11-21.
- [30] Gantikota, S. (2023). Integrating SonarQube and IBM AppScan into Enterprise CI/CD Pipelines: A Vulnerability Mitigation Framework Achieving Over Eighty Percent Risk Reduction. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(3), 240-244. <https://doi.org/10.63282/3050-9246.IJETCSIT-V4I3P124>
- [31] Brahmandam, L. M. K. (2024). An Empirical Evaluation of the Medallion Architecture on Databricks and Apache Spark with Snowflake: Throughput, Latency, and Cost for Batch and Real-Time Ingestion Patterns. *International Journal of AI, BigData, Computational and Management Studies*, 5(3), 197-206. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V5I3P122>
- [32] Nangi, P. R., Obannagari, C. K. R. N., & Settipi, S. (2025). Large Language Model-Driven Intelligent Observability Frameworks for Serverless Applications in Event-Driven Cloud Architectures. *International Journal of Emerging Research in Engineering and Technology*, 6(1), 100-110.