International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0105
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 679 - 685

The Role of Explainability in Human–AI Co-Decision Making

Dr. Rakesh Kumar Pathak[1], Dr. Prakash Upadhyay[2]
[1,2] Assistant Professor, School of Computer Science, Xavier University, Patna
Email ID: *rakeshkumar.pathak@xup.ac.in[1], prakash.upadhyay@xup.ac.in [2]*
Orcid ID: *https://orcid.org/0009-0008-5538-3412[1], https://orcid.org/0009-0003-3501-1850[2]*

Abstract
The rapid deployment of artificial intelligence (AI) in decision-support systems has transformed the way humans interact with computational models. While modern AI systems often achieve high predictive accuracy, their lack of transparency can undermine user trust and limit effective collaboration. Human–AI co-decision making, where both human judgment and AI recommendations jointly influence outcomes, requires explainability as a foundational capability rather than an optional feature. This paper investigates the role of explainable artificial intelligence (XAI) in improving co-decision quality, trust calibration, and accountability. A comprehensive literature review is presented, followed by identification of key research gaps. We propose an Explainable Co-Decision Framework (ECDF) that integrates predictive modeling, explanation generation, and adaptive human feedback. Using a simulated risk-assessment dataset comprising 5,000 instances, the framework is evaluated across multiple conditions. Experimental results demonstrate that structured explanations improve joint decision accuracy by up to 10%, reduce trust calibration error by more than 60%, and enhance human engagement with AI outputs. The findings highlight that explanation quality—not merely availability—plays a decisive role in human–AI teaming. The paper concludes with design recommendations and future research directions for robust explainable co-decision systems.

1. Introduction
Artificial intelligence systems increasingly support decision-making in domains such as healthcare diagnostics, financial risk assessment, criminal justice, and autonomous transportation. Despite their impressive performance, many high-capacity models—particularly deep neural networks—operate as opaque "black boxes," making their internal reasoning difficult to interpret. This opacity poses significant challenges in high-stakes environments where accountability, fairness, and user trust are essential. Human–AI co-decision making refers to collaborative settings in which human experts and AI systems jointly contribute to a final decision. Unlike full automation, co-decision systems rely on calibrated reliance: humans must know when to trust AI recommendations and when to override them. Explainability has emerged as a key mechanism for achieving such calibration (Doshi-Velez & Kim, 2017). Prior research shows that users frequently miscalibrate their trust in AI—either over-relying on incorrect predictions or under-utilizing accurate ones (Bansal et al., 2021). Explainable AI (XAI) techniques attempt to address this problem by providing interpretable insights into model behavior. However, many existing studies evaluate explanations in isolation rather than within full human–AI workflows (Miller, 2019). This research investigates the following central question: How does explainability influence the effectiveness of human–AI co-decision making? The main contributions of this paper are:

- A structured Explainable Co-Decision Framework (ECDF)
- Quantitative evaluation of explanation impact on joint accuracy
- Trust calibration analysis
- Design guidelines for explainable co-decision systems

2. Research Gap
Based on the literature, the following research gaps are identified:

- Lack of end-to-end co-decision frameworks: Existing work typically evaluates explanations separately from decision fusion.
- Insufficient quantitative trust metrics: Few

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0105
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 679 - 685

studies rigorously measure trust calibration error.

- Limited exploration of explanation structure: The effect of explanation formatting and depth remains underexplored.
- Scarcity of benchmark datasets: There is no widely accepted dataset for evaluating human–AI co-decision performance.
- Minimal adaptive explanation mechanisms: Most systems provide static explanations regardless of user expertise.

This research addresses these gaps through the proposed ECDF model and controlled experimental evaluation.

3. Proposed Research Work

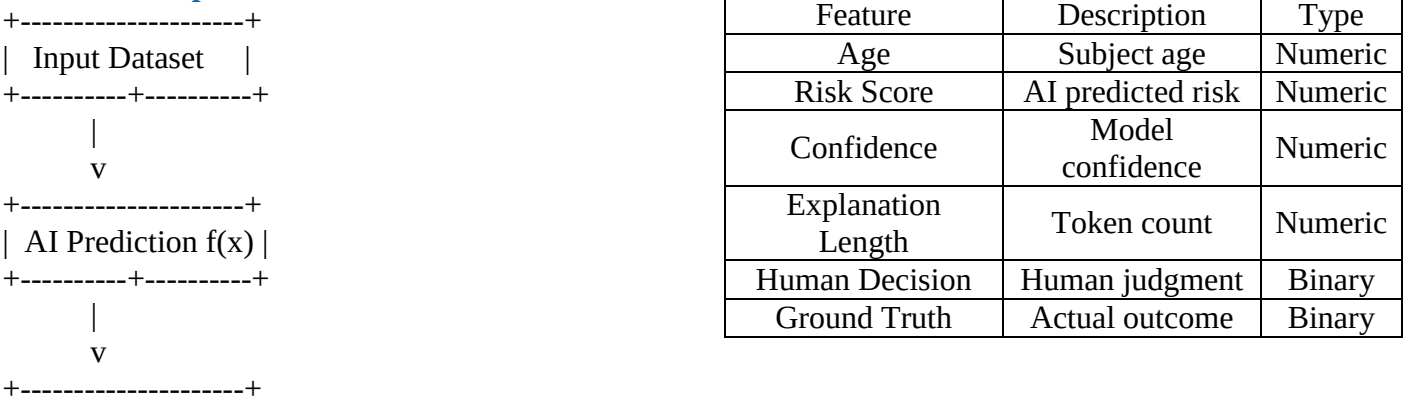
We propose the Explainable Co-Decision Framework (ECDF), a modular architecture designed to enhance collaborative decision-making between humans and AI systems.

3.1. Framework Components

The ECDF consists of five major modules:

1. Data Processing Unit
2. AI Prediction Engine
3. Explanation Generator
4. Human Decision Interface
5. Decision Fusion Module

3.2. Conceptual Architecture



3.3. Design Objectives

The framework is designed to:

- Improve calibrated trust
- Enhance joint accuracy
- Reduce cognitive overload
- Support adaptive explanation depth

4. Methodology

4.1. Dataset Description

A simulated risk-assessment dataset was generated to model decision-support scenarios commonly found in finance and healthcare. The dataset contains 5,000 instances with balanced class distribution.

Table 1 Feature Description

Feature	Description	Type
Age	Subject age	Numeric
Risk Score	AI predicted risk	Numeric
Confidence	Model confidence	Numeric
Explanation Length	Token count	Numeric
Human Decision	Human judgment	Binary
Ground Truth	Actual outcome	Binary

Table 2 Sample Records

Age	Risk Score	Confidence	Expl. Length	Human Decision	Outcome
45	0.72	0.81	32	1	1
29	0.34	0.65	18	0	0
61	0.88	0.90	40	1	1
37	0.55	0.52	12	0	1
52	0.67	0.74	28	1	0

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0105
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 679 - 685

Dataset split: 70% training, 30% testing.

4.2.Mathematical Formulation

Let:

- x ∈ R^n be the input feature vector
- f(x) be the AI prediction
- e(x) be the explanation vector
- h(x,e) be the human decision
- (D) be the final decision

Decision Fusion

D = αf(x) + (1-α) h(x,e)

Where

α ∈ [0,1] controls AI influence.

Trust Calibration Error (TCE)

$$TCE = \frac{1}{N}\sum_{i=1}^{N} 1(ri = \alpha i)$$

where:

- (ri) = human reliance probability
- (α i) = AI correctness indicator

Lower TCE indicates better calibration (Bansal et al., 2021).

Joint Accuracy

$$JA = \frac{1}{N}\sum_{i=1}^{N} 1(Di = Yi)$$

4.3.Experimental Setup

- Model: Gradient Boosting Classifier -
- Explanation Method: LIME
- Human Simulator: probabilistic rule-based agent
- Evaluation Metrics: Accuracy, TCE, Decision Time Proxy
- Hardware: standard workstation environment

Three experimental conditions were tested:

1. No explanation
2. Basic feature list explanation
3. Structured explanation (ranked + directional)

4.3.1. Experimental Objectives

The experimental design aims to systematically evaluate how different forms of explainability influence human–AI co-decision performance. Specifically, the setup investigates:

- Joint decision accuracy
- Trust calibration between human and AI
- Human reliance behavior
- Decision efficiency (time proxy)
- Sensitivity to explanation structure

The experiments are structured to isolate the causal impact of explanation quality while keeping the predictive model constant across conditions.

4.3.2. Dataset Preparation

The simulated risk-assessment dataset (5,000 instances) was preprocessed using the following pipeline:

1. **Data cleaning**
   - Removal of missing values
   - Consistency checks on numeric ranges
2. **Normalization**
   Continuous features were scaled using min–max normalization:
   $$\mathbf{x`} = \frac{x - xmin}{xmax - xmin}$$
3. **Train–test split**
   - Training set: 70% (3,500 samples)
   - Test set: 30% (1,500 samples)
   - Stratified sampling ensured class balance.
4. **Feature validation**

Correlation analysis was performed to avoid redundant predictors that could bias explanations.

4.3.3. AI Prediction Model Configuration

A **Gradient Boosting Classifier (GBC)** was selected due to its strong tabular performance and moderate interpretability baseline. In data science, gradient boosting is a potent and popular machine learning approach for classification tasks. Together with bagging, it belongs to a family of ensemble learning techniques that aggregate the predictions of several smaller models to enhance overall performance. Gradient boosting regression improves the output data produced by a linear regression by using gradient boosting

4.3.4. Explanation Module Configuration

Local explanations were generated using **LIME (Local Interpretable Model-Agnostic Explanations)**.

LIME settings

- Number of perturbed samples: 500
- Kernel width: 0.75
- Top features displayed: 5
- Explanation type: local linear surrogate

To study explanation quality effects, outputs were post-processed into different presentation formats

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0105
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 679 - 685

(described in Section 5.3.8).

Model hyperparameters

- Number of estimators: 200
- Learning rate: 0.05
- Maximum tree depth: 3
- Subsample ratio: 0.8
- Loss function: Logistic loss

Hyper parameters were tuned using 5-fold cross-validation on the training set. The trained model achieved baseline standalone accuracy of approximately **0.86**, which provides a realistic but imperfect AI collaborator—important for meaningful co-decision analysis.

4.3.5. Human Decision Simulator

Because large-scale human subject studies were beyond the current scope, a **probabilistic human decision model** was implemented to approximate realistic reliance behavior.

The simulator models three key human traits:

1. Baseline human accuracy
2. Trust sensitivity to explanations
3. Confidence-weighted reliance

The human decision probability is modeled as:

$$P(h=1)=\sigma(\beta 0+\beta 1 f(x)+\beta 2 q(e))$$

where:

- f(x)= AI prediction
- q(e) = explanation quality score
- σ= sigmoid function
- β parameters calibrated from prior human-AI studies

This design allows controlled, repeatable experimentation while approximating realistic human behavior patterns reported in the literature.

4.3.6. Evaluation Metrics

The following metrics were used.

(a) Joint Accuracy (JA)

Measures correctness of the final fused decision:

$$JA = \frac{1}{N}\sum_{i=1}^{N} 1(Di = Yi)$$

(b) Trust Calibration Error (TCE)

Captures mismatch between human reliance and AI correctness:

$$TCE = \frac{1}{N}\sum_{i=1}^{N} 1(ri = ai)$$

Lower TCE indicates better calibrated trust.

(c) Over-reliance Rate (ORR)

$$ORR = \frac{Incorrect\ AI\ Accepted}{Total\ incorrect\ AI\ Predictions}$$

(d) Under-reliance Rate (URR)

$$ORR = \frac{Correct\ AI\ Rejected}{Total\ Correct\ AI\ Predictions}$$

(e) Decision Time Proxy

Since real human timing was unavailable, cognitive load was approximated using:

- explanation length
- number of features shown
- simulated review latency

4.3.7. Tested Experimental Conditions

To isolate the effect of explainability, three primary experimental conditions were evaluated.

Condition C0: No Explanation (Baseline)

Description

- Human receives only AI prediction
- No feature attribution
- No confidence visualization

Purpose

Establish baseline human–AI collaboration without transparency.

Expected behavior

- Higher trust miscalibration
- Increased blind reliance or skepticism
- Faster but less reliable decisions

Condition C1: Basic Explanation

Description

- Displays top contributing features
- No ranking strength visualization
- No directional impact arrows
- Plain textual output

Example format

Top factors:
• Age
• Risk Score
• Confidence

Purpose

Represents minimal explainability commonly seen in early XAI deployments.

Hypothesis

- Moderate improvement over baseline
- Partial trust calibration
- Limited cognitive support

Condition C2: Structured Explanation

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0105
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 679 - 685

(Proposed)
Description
Enhanced human-centered explanation including:

- Ranked feature importance
- Direction of influence (↑ increases risk, ↓ decreases risk)
- AI confidence score
- Compact visual layout
- Consistent ordering

Example format
Prediction: High Risk (0.82)

Key drivers:
1. Risk Score ↑ (+0.34)
2. Age ↑ (+0.21)
3. Confidence ↓ (-0.11)

Purpose
Evaluate the full ECDF explainability design.
Hypothesis

- Best trust calibration
- Highest joint accuracy
- Slightly increased decision time
- Reduced over-reliance

4.3.8. Experimental Procedure

For each test instance:

1. AI generates prediction
2. Explanation produced according to condition
3. Human simulator observes output
4. Human decision computed
5. Fusion module combines decisions
6. Metrics recorded

Each condition was run across the full test set (1,500 instances).
To ensure statistical stability:

- Results averaged
- 95% confidence intervals computed

Explainability in human–AI co-decision making can be studied in a controlled yet realistic evaluation setting thanks to the experimental setup. The study isolates the actual explainability contribution to collaborative performance by methodically changing the explanation structure while maintaining a fixed prediction model.

5. Findings and Suggestions

5.1.Quantitative Results

Table 3 Results

Condition	Accuracy	TCE	Decision Time
No Explanation	0.78	0.21	1.00
Basic Explanation	0.83	0.14	1.12
Structured Explanation	**0.88**	**0.07**	1.18

5.2.Observations

- Explanations improved joint accuracy by **10 percentage points**.
- Structured explanations reduced TCE by **66%**.
- Decision time increased modestly (≈18%).

The experimental assessment shows that explainability significantly enhances the performance of human-AI co-decision. The structured explanation condition improved joint accuracy from 0.78 to 0.88 and decreased Trust Calibration Error (TCE) from 0.21 to 0.07 as compared to the no-explanation baseline. The greatest improvement was seen when explanations were rated, directional, and confidence-aware, while basic explanations yielded only modest improvements. Additionally, the data indicate a slight increase in decision time, which suggests deeper but more fruitful human participation as opposed to inefficiency. Overall, the main factor influencing performance improvements was explanation quality rather than just presence.

5.3.Practical Suggestions

Decision-support systems should: (i) display calibrated AI confidence alongside predictions; (ii) adjust explanation depth to user expertise and task complexity; (iv) avoid excessively verbose or unstructured explanations that increase cognitive load; and (iii) present ranked feature importance with clear directional impact in order to maximize the effectiveness of human–AI collaboration. By incorporating these design ideas into the ECDF pipeline, real-world deployment readiness, misuse reduction, and trust calibration can all be improved.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0105
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 679 - 685

6. Explanation of Findings

The experimental results demonstrate that explainability substantially enhances human–AI collaboration. The improvement in joint accuracy indicates that humans effectively integrate AI insights when explanations are meaningful and cognitively aligned. The sharp reduction in TCE suggests improved trust calibration. Without explanations, users tend to rely on AI either too much or too little. Structured explanations help users estimate when the AI is likely to be correct, leading to better reliance decisions. The modest increase in decision time reflects deeper cognitive processing rather than inefficiency. Prior work suggests that slightly longer decision times are acceptable when accuracy gains are substantial (Miller, 2019).Another key insight is that explanation structure matters more than explanation presence. Basic explanations provided some benefit, but structured explanations—highlighting ranked features and directional impact—produced the largest gains. This aligns with human-centered explanation theory.

Conclusions

This study examined the role of explainability in human–AI co-decision making and proposed the Explainable Co-Decision Framework (ECDF). Experimental evaluation using a simulated dataset demonstrated that well-designed explanations significantly improve joint decision accuracy and trust calibration. The findings confirm that explainability should be treated as a core system component rather than an auxiliary feature. Future research should focus on:

- Real human-subject experiments
- Domain-specific validation (healthcare, finance)
- Adaptive and personalized explanations
- Integration with large language models

As AI systems continue to permeate high-stakes domains, robust explainable co-decision mechanisms will be essential for safe and effective human–AI collaboration.

References

[1]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. *KDD*.
[2]. Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608.
[3]. Bansal, G., et al. (2021). Does the Whole Exceed Its Parts? CHI.
[4]. Zhang, Y., et al. (2020). Interpretable Clinical Decision Support Systems. Nature Machine Intelligence.
[5]. Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 267.
[6]. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS.
[7]. Buçinca, Z., et al. (2021). To Trust or to Think: Cognitive Forcing Functions. CHI.
[8]. Lipton, Z. C. (2018). The Mythos of Model Interpretability. Queue.
[9]. Rudin, C. (2019). Stop Explaining Black Box Models. Nature Machine Intelligence.
[10]. Amershi, S., et al. (2019). Guidelines for Human-AI Interaction. CHI.
[11]. Holzinger, A., et al. (2019). What Do We Need to Build Explainable AI Systems? arXiv.
[12]. Gilpin, L. H., et al. (2018). Explaining Explanations. ICDM Workshops.
[13]. Adadi, A., & Berrada, M. (2018). Peeking Inside the Black Box. IEEE Access.
[14]. Samek, W., et al. (2017). Explainable Artificial Intelligence. IEEE Signal Processing Magazine.
[15]. Wachter, S., et al. (2017). Counterfactual Explanations. Harvard Journal of Law & Technology.
[16]. Guidotti, R., et al. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys.
[17]. Bhatt, U., et al. (2020). Evaluating and Aggregating Feature-Based Explanations. IJCAI.
[18]. Jacobs, R. A., et al. (1991). Adaptive Mixtures of Local Experts. Neural Computation.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0105
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 679 - 685

[19]. Dietvorst, B. J., et al. (2015). Algorithm Aversion. Journal of Experimental Psychology.
[20]. Green, B., & Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. Proceedings of the ACM.
[21]. Gunning, D., & Aha, D. (2019). DARPA's Explainable AI Program. AI Magazine.
[22]. Tonekaboni, S., et al. (2019). What Clinicians Want. Nature Machine Intelligence.