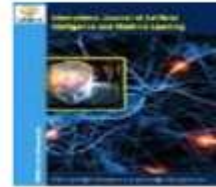




International Journal of Artificial
Intelligence and Machine Learning
Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Explainable Artificial Intelligence in Human–AI Collaboration: Enhancing Trust and Interpretability in Decision Systems

Shatarupa Sarma

AngleIQ Inc, USA. ORCID: 0009-0001-6695-9439

Abstract

Explainable artificial intelligence (XAI) addresses the challenge of understanding, validating, and appropriately trusting model outputs in consequential decision environments. This article introduces a formal evaluation framework comprising three quantitative measures for assessing XAI effectiveness in human–AI collaborative decision systems: the Trust Calibration Score (TCS), which quantifies alignment between user trust and actual system accuracy; the Explanation Utility Index (EUI), which combines comprehension gain and decision quality improvement in a weighted composite; and the Human–AI Collaboration Gain (HACG), which measures performance surplus or deficit relative to individual baselines. Utilizing existing XAI methods (LIME, SHAP, Grad-CAM, and counterfactuals), the framework is applied across three deployment conditions and three application scenarios. XAI-supported conditions yield TCS values above 0.95, compared to 0.74 under the black-box condition, with HACG results showing a 7–9% improvement over the strongest individual baseline. The proposed framework provides a replicable methodology for evaluating and optimizing XAI implementations across clinical, financial, and organizational decision environments.

Keywords: Explainable Artificial Intelligence, Trust Calibration, Human-AI Collaboration, Interpretability, SHAP, LIME, Decision Support Systems

1. Introduction

The integration of artificial intelligence into consequential decision processes, including clinical diagnosis, financial risk assessment, legal sentencing support, and industrial safety monitoring, has created urgent demands for transparency and accountability in model behavior [5, 7, 9]. While contemporary AI systems achieve high predictive accuracy across many domains, precision alone does not guarantee effective human-AI collaboration. Decision makers who cannot understand how a model reaches its conclusions are unable to verify its outputs, integrate domain knowledge appropriately, or detect systematic model errors before they propagate to harmful decisions.

A central manifestation of this challenge is trust miscalibration: the divergence between a user's subjective trust in an AI system and the system's actual performance. Over-trust occurs when humans uncritically accept the model's output and accept it as correct, overlooking mistakes that the human would be able to detect [8, 11]. Under-trust occurs when humans are not aware of, dismiss, or reject correct AI recommendations (e.g., rejecting the output of a recommender system). These two failure modes degrade the quality of human-AI collaborative decision-making and weaken the case for the introduction of AI into high-stakes environments [10, 17].

The field of explainable artificial intelligence (XAI) seeks to remedy the miscalibration of trust by providing human-interpretable explanations of model decisions [1, 7]. Foundational methods including LIME [2], SHAP [3], Grad-CAM [12], and counterfactual explanations [13] provide different explanation modalities suited to different model types, data domains, and user needs. Comprehensive surveys [4, 6] have catalogued the properties and trade-offs of these methods, and application studies [15, 18, 20] have documented their impact across clinical, financial, and organizational domains. Despite this progress, the field lacks a unified quantitative framework for evaluating XAI effectiveness across the dimensions most critical to human-AI collaboration: trust calibration, explanation utility, and collaborative performance gain.

This article addresses that gap by introducing three formal measures. Another metric to evaluate explainability is the Trust Calibration Score (TCS), a quantitative measure of how well subjective trust matches the objectively measured performance of a system. The Explanation Utility Index (EUI) consolidates users' increase in understanding and decision-making into a composite score weighted by their relative importance. The Human-

AI Collaboration Gain (HACG) measures the performance surplus or deficit achieved by a human-AI team relative to the strongest individual baseline. Together, these measures form a compact diagnostic framework for evaluating, benchmarking, and optimizing XAI implementations.

Specifically, this article makes the following contributions:

1. A formal trust calibration metric (TCS) that quantifies the alignment between user trust and system accuracy, providing a symmetric measure applicable across both over-trust and under-trust conditions.
2. A weighted composite metric (EUI) that integrates comprehension gain and decision quality improvement, adaptable to domain-specific priorities through configurable weights.
3. A collaboration performance metric (HACG) that isolates the value added by human-AI teaming relative to either party operating independently, enabling direct comparison of XAI and non-XAI deployment conditions.
4. Illustrative numerical analyses across three deployment conditions and three application domains demonstrating the diagnostic application of the TCS-EUI-HACG framework

2. Background and Related Work

2.1 Foundations of XAI and Interpretability

The scientific study of interpretable machine learning was formalized by Doshi-Velez and Kim [1], who argued that interpretability must be evaluated empirically against human task performance rather than defined abstractly. Lipton [5] identified a persistent conceptual ambiguity in the field: interpretability is invoked simultaneously to mean simulatability, decomposability, and algorithmic transparency, properties that do not always coincide in practice. Murdoch et al. [6] provided a definitional framework distinguishing descriptive, scientific, and predictive accuracy as distinct goals, each requiring different interpretability properties.

XAI research was institutionalized under the DARPA Explainable AI Program [9], which funded various model types and explanation methods with an explicit goal of maintaining high task performance while increasing human comprehension. Molnar et al. [4] surveyed the resulting landscape, categorizing methods by scope (local versus global), type (model-agnostic versus model-specific), and output format (feature attributions, rules, exemplars, and counterfactuals).

2.2 Trust in Human-AI Systems

Trust in AI systems is a dynamic calibration process that responds to evidence about system accuracy, transparency, and past behavior [8]. Miller [8] argues that explanations are social acts whose effectiveness depends on the recipient's beliefs, cognitive load, and motivational context, necessitating user-centered design. Wang et al. [11] found that model competence and explanation quality have independent effects on acceptance of AI recommendations.

Naiseh et al. [10] studied trust calibration in AI-based recommender systems and found a systematic effect of model-agnostic explanation type and detail on reliance behavior: explanations with model uncertainty reduced over-reliance more strongly than explanations with only feature contributions communicated. de Brito Duarte et al. [17] investigated whether XAI can generate warranted trust, finding that explanation-enhanced systems supported more accurate trust calibration but that the benefit varied substantially across user expertise levels, a finding that motivates the user-stratified analysis presented in Section 6.

2.3 Interpretability Methods

LIME [2] generates local post-hoc explanations by fitting an interpretable surrogate model to the neighborhood of a specific prediction, producing feature weights that indicate each feature's contribution to the prediction. SHAP [3] computes feature attributions using Shapley values from cooperative game theory, assigning each feature the average marginal contribution it makes across all possible feature subsets. SHAP satisfies a rigorous set of theoretical axioms, including local accuracy, missingness, and consistency, that guarantee certain desirable properties of the resulting attributions. Consequently, Salih et al. [14] found that for a series of classification tasks, SHAP attributions were more consistent than LIME attributions, at the cost of extremely high computation time, a serious issue for real-time decision support.

Grad-CAM [12] highlights salient image regions that contribute most to the model's prediction, providing a visualized localization of supported predictions in the form of gradient-weighted activation maps. Counterfactual explanations [13] display the changes that would be necessary for the model's prediction to differ; this approach aligns well with the GDPR's right to explanation. Kaur et al. [16] studied how data scientists interpret these explanation methods in the wild and found that even technically savvy users systematically misinterpret SHAP outputs. Xu et al. [19] reviewed human-AI collaborative systems for non-experts using a model of intuitive decision-making and found that the design of an explanation mediates the relationship between an AI output and a user's decision-making quality.

2.4 Research Gap and Motivation

Despite the substantial body of work on individual XAI methods and their effects on trust and performance, a critical gap persists: no existing framework integrates trust calibration, explanation utility, and collaborative

performance gain into a unified evaluation methodology. Current approaches evaluate these dimensions in isolation, which limits the ability to diagnose the root cause of suboptimal human–AI collaboration. Existing trust metrics do not account for explanation quality, comprehension measures do not link to collaborative outcomes, and performance evaluations rarely isolate the contribution of the explanation layer from the model itself. Moreover, most existing evaluation methods are tied to specific domains or explanation types, lacking the generalizability needed for cross-domain benchmarking. The framework proposed in this article addresses these limitations by providing three complementary, domain-adaptable metrics that together enable a comprehensive diagnostic assessment of XAI-enabled systems. The following sections formalize the XAI methods used (Section 3) and the proposed evaluation metrics (Section 4).

3. XAI Methods and System Architecture

3.1 Post-hoc Local Explanation Methods

Post-hoc explanation methods generate interpretable accounts of model behavior after the model has produced a prediction, without requiring access to the model's training process or internal weights. This model-agnostic property makes them applicable to any classifier, including proprietary or commercially deployed systems [2]. LIME generates a local explanation by sampling perturbations of the input, querying the model for predictions on each perturbation, and fitting a locally linear model to the resulting prediction surface. The linear model's coefficients serve as the explanation: features with large positive coefficients increased the predicted class probability, while features with large negative coefficients decreased it. A key strength of LIME is its computational efficiency and intuitive output format; however, the perturbation-based approach introduces variance in explanations across repeated runs, and the linear approximation may poorly represent nonlinear decision boundaries.

SHAP extends this post-hoc approach by grounding feature attributions in cooperative game theory [3]. However, exact SHAP computation is NP-hard, and kernel-based approximations introduce their own variance, particularly in high-dimensional feature spaces. Each feature receives an attribution equal to its average marginal contribution across all possible subsets of features, computed over a background reference distribution. Unlike LIME, SHAP provides global consistency guarantees: if a model changes such that a feature becomes more important, the SHAP value for that feature will not decrease. This consistency property is particularly valuable in regulated environments where explanation stability across similar inputs is a compliance requirement. Table 1 summarizes the principal XAI methods discussed in this article.

Table 1. XAI Method Comparison: Type, Model-Agnosticism, Output Format, and Use Case

Method	Explanation Type	Model Agnostic	Output Format	Primary Use Case	Key Limitations
LIME	Local post-hoc	Yes	Feature weights	Per-instance classification explanation	Perturbation variance; local linearity assumption
SHAP	Local post-hoc	Yes	Shapley values	Consistent feature attribution across tabular, text, and image data	High computation cost; kernel approximation variance
Grad-CAM	Gradient-based visual	No	Saliency map	Spatial localization in convolutional neural networks	CNN-specific; coarse resolution at deeper layers
Counterfactual	Contrastive	Yes	Minimal change set	Actionable recourse and regulatory compliance (GDPR)	Multiple valid counterfactuals; actionability constraints
Interpretable Models (e.g., DT, LR)	Intrinsic	N/A	Model structure	High-transparency applications requiring auditable logic	Limited model complexity and accuracy

3.2 Gradient-Based and Visual Explanation Methods

Grad-CAM generates visual explanations for convolutional neural network predictions by computing the gradient of the target class score with respect to the final convolutional layer's feature maps [12]. These gradients are globally average-pooled to produce a weight vector representing each feature map's importance for the target class. This weighted combination of feature maps is then passed through a ReLU to produce the final saliency map of the input image, which shows which areas were most important for a given network decision. Thus, saliency maps provide an intuitive spatial explanation of what was important to the network without access to intrinsic model parameters beyond computing gradients.

3.3 Counterfactual Explanations

Counterfactual explanations identify the minimum change to an input that would produce a different model output [13]. They answer the question: what would need to be different about this case for the decision to change? This form of explanation is particularly aligned with human causal reasoning and is directly actionable: a loan applicant who receives a counterfactual explanation knows which attributes to change to receive a positive outcome in the future. Wachter et al. [13] demonstrated that counterfactuals can be generated without access to model internals, making them practical for explaining opaque commercial systems in regulatory contexts.

3.4 XAI System Architecture

An XAI-enabled decision system integrates an explanation layer between the model output and the human decision interface. The architecture follows a five-stage pipeline: raw data is processed and transformed into model-compatible inputs; the model generates a prediction with an associated confidence score; the explanation layer applies the selected XAI method to generate a human-interpretable account of the prediction; the explanation is rendered in a format matched to the user's expertise and the decision context; and the human decision maker uses both the prediction and the explanation to form a final decision or validation judgment. This pipeline directly supports the proposed evaluation metrics: TCS is measured at stage 5 (comparing user trust to system accuracy), EUI is measured across stages 3 through 5 (explanation quality and its effect on comprehension and decisions), and HACG is measured by comparing stage 5 outcomes across collaborative and individual conditions.

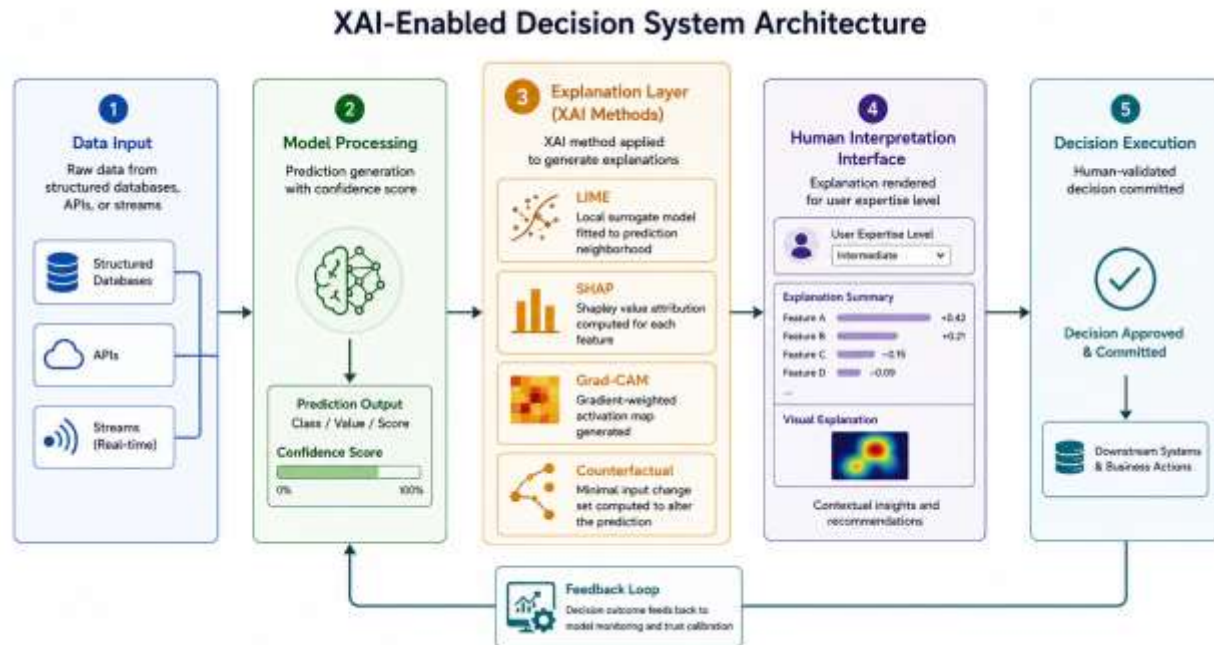


Figure 1: XAI-Enabled Decision System Architecture

4. Mathematical Framework for XAI Evaluation

Evaluating XAI effectiveness requires measures that capture distinct aspects of the human-AI interaction: how accurately user trust reflects system performance, how much explanations improve comprehension and decision quality, and whether human-AI collaboration produces better outcomes than either party alone. The three metrics proposed below were designed to satisfy four criteria: (a) formal precision sufficient for replicable computation, (b) interpretability via defined threshold zones, (c) adaptability to domain-specific priorities through configurable parameters, and (d) complementarity, such that the three metrics together diagnose failure modes that any single metric would miss. Existing evaluation approaches in the XAI literature typically address only one of these dimensions. Trust surveys [10, 17] measure calibration but not explanation quality; task accuracy comparisons [11] measure performance but not the mechanism by which XAI contributes; and comprehension assessments [16] measure understanding but not its downstream effect on decisions. The TCS-EUI-HACG framework integrates these dimensions into a single diagnostic methodology.

4.1 Trust Calibration Score (TCS)

Trust Calibration Score (TCS) quantifies the degree to which users' subjective trust in an AI system is consistent with the system's actual or objective performance:

Formula 1 (TCS): $TCS = 1 - |T_u - A_s|$

The user trust score T_u has a range of $[0, 1]$ and is often calculated through surveys or through proxies such as the acceptance rate of the AI recommendation. The system accuracy, denoted as A_s , represents the true accuracy of the system on a representative task, normalized to the range $[0, 1]$.

TCS ranges from 0, representing maximum miscalibration, to 1, representing perfect calibration. A TCS value of 0.80 or above indicates well-calibrated trust: the system closely aligns with the user's subjective confidence in its actual performance. A TCS between 0.60 and 0.79 indicates moderate miscalibration, where the divergence between trust and accuracy is detectable and may influence decision quality. A TCS below 0.60 indicates severe miscalibration requiring intervention, such as targeted explanation design or user training. The measure captures miscalibration symmetrically: it does not distinguish between over-trust, where T_u exceeds A_s , and under-trust, where T_u falls below A_s , as both states are associated with suboptimal collaborative performance [10, 11].

Worked example. A clinical decision support system has a verified accuracy of $A_s = 0.85$. A clinician survey yields $T_u = 0.83$. Then $TCS = 1 - |0.83 - 0.85| = 1 - 0.02 = 0.98$, indicating well-calibrated trust.

Assumptions and limitations. TCS assumes that trust and accuracy can be meaningfully measured on commensurate $[0, 1]$ scales, that a single scalar trust value adequately summarizes the user's trust state, and that symmetric treatment of over-trust and under-trust is appropriate for the application. In domains where over-trust carries substantially higher risk than under-trust (e.g., autonomous driving), an asymmetric variant with directional penalties may be warranted.

4.2 Explanation Utility Index (EUI)

The Explanation Utility Index assesses the functional value of an explanation to a human user by combining two measurable outcomes of explanation exposure: comprehension gain and decision quality improvement. It is defined as:

Formula 2 (EUI): $EUI = (\Delta C \times w_c + \Delta D \times w_d) / (w_c + w_d)$

Where ΔC is the comprehension gain score, computed as the difference between post-explanation and pre-explanation comprehension assessments, normalized to $[0, 1]$; ΔD is the score for improvement in decision quality, computed as the difference in task accuracy between conditions with and without explanation, normalized to $[0, 1]$; and w^c and w^d are domain-specific weights that sum to 1, reflecting the relative importance of understanding versus performance in the target application context.

EUI values are interpreted according to three utility zones. A value of 0.70 or above denotes high utility: the explanation substantially improves both user understanding and decision quality. 0.40 to 0.69 means moderately useful: the explanation produces meaningful, but not fully complete, improvements. A value below 0.40 denotes low utility, indicating that the explanation modality, format, or complexity is poorly matched to the user's needs or expertise level and that redesign is warranted. The weighting parameters w^c and w^d allow EUI to be adapted to application-specific priorities: in clinical decision support [15], decision quality typically receives higher weight; in audit and compliance contexts, comprehension may be prioritized.

Worked example. A fraud detection system provides LIME explanations to non-expert analysts. Pre-explanation comprehension: 0.40; post-explanation: 0.85; $\Delta C = 0.45$. Task accuracy without explanation: 0.72; with explanation: 1.10; $\Delta D = 0.38$. With equal weights ($w^c = w^d = 0.50$): $EUI = (0.45 \times 0.50 + 0.38 \times 0.50) / 1.0 = 0.415$, indicating moderate utility.

Assumptions and limitations. EUI assumes that comprehension and decision quality can be independently measured, that the weighted linear combination is an adequate aggregation, and that the weight parameters are set by informed domain judgment. The metric does not capture explanation side effects such as increased cognitive load or decision time, which may be relevant in time-critical environments.

4.3 Human-AI Collaboration Gain (HACG)

The Human-AI Collaboration Gain measures whether a human-AI team achieves better performance than the best individual baseline. It is defined as:

Formula 3 (HACG): $HACG = (P_{collab} - \max(P_{human}, P_{AI})) / \max(P_{human}, P_{AI})$

Where P_{collab} is the performance score achieved in a human-AI collaborative condition; P_{human} is the performance score in a human-only baseline condition; and P_{AI} is the performance score in an AI-only baseline condition.

HACG takes a positive value when the collaborative team outperforms the best individual, indicating complementary gains from the human-AI interaction. A human with a positive HACG has knowledge of the domain, can identify errors made by the model, and can make use of contextual information not available to the AI model itself because the reasoning is transparent through the explanations. A HACG of 0 means the team size is indistinguishable from the best individual. Negative HACGs indicate that a team lacks collaboration, performing worse than the best individual. They have been attributed to miscalibrated trust and poorly designed explanations [11, 17]. Without XAI, a user may accept an AI's output as is, leading to the same pattern of mistakes from both the AI and the user, but without the user correcting the AI.

Worked example. In a document review task without XAI, $P_{\text{human}} = 0.76$, $P_{\text{AI}} = 0.82$, and $P_{\text{collab}} = 0.79$. Then $\text{HACG} = (0.79 - 0.82) / 0.82 = -0.037$, a 3.7% collaboration deficit caused by automation bias.

Assumptions and limitations. HACG assumes that performance can be measured on a common scale across human, AI, and collaborative conditions. The metric normalizes against the best individual, so small differences near the baseline can produce volatile HACG values. HACG also does not decompose the sources of collaboration gain or deficit; a positive HACG could result from error correction, complementary coverage, or other mechanisms, which would need to be investigated through finer-grained analysis

5. Human-AI Interaction and Trust Dynamics

5.1 Trust Calibration in Collaborative Systems

Trust calibration in human-AI systems is determined by the interplay of three factors: the quality and stability of the AI output, the transparency with which the AI explains its behavior, and the user's prior knowledge and level of expertise [8, 10]. When users integrate sufficient evidence about what the AI can and cannot do, they can better calibrate their trust. XAI addresses this by giving structured information on the features and context that resulted in a particular model's prediction, allowing users to compare their understanding with that of the model and update their trust accordingly.

Naiseh et al. [10] found that explanation designs that surface model uncertainty, such as SHAP confidence intervals or counterfactual proximity scores, produce more accurate trust calibration than explanation designs that present only point estimates of feature importance. This finding suggests that TCS can be improved not only by increasing explanation accuracy but also by changing what aspects of model behavior the explanation makes visible. de Brito Duarte et al. [17] confirmed that well-designed XAI generates warranted trust, a form of trust that is sensitive to actual model performance rather than stylistic features of the interface.

5.2 Explanation Design for Diverse User Populations

The utility of an explanation, as measured by EUI, is not an intrinsic property of the explanation method but depends substantially on the expertise and cognitive profile of the intended user [4, 16]. Kaur et al. [16] documented systematic misinterpretation of SHAP values among technically trained data scientists, finding that users frequently overgeneralized local feature attributions to global model behavior. However, evidence suggests that non-expert users have different anchoring patterns, favoring the most dominant explanation over the full distribution of attribution possibilities.

These observations suggest designing explanations for different levels of expertise. Domain experts can benefit from the granularity of feature attribution methods like SHAP that can align AI reasoning with human domain expertise. For non-expert users, visualizations or natural language summaries of the feature attributions may be more effective. Counterfactual explanations [13] may be more effective because they show how a feature is relevant to what the user must change. Xu et al. [19] reviewed evidence for an intuitive decision-making model for human-AI collaboration and found that matching the user's default reasoning mode resulted in higher EUI values.

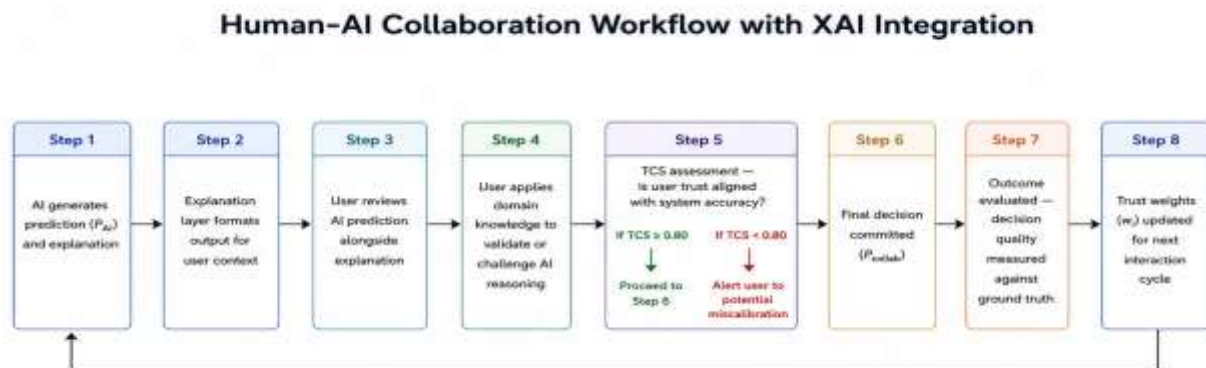


Figure 2: Human-AI Collaboration Workflow with XAI Integration

5.3 Cognitive Factors in Explanation Acceptance

Cognitive factors such as cognitive load, confirmation bias, and automation bias are known to moderate the relationship between exposure and use of explanations [8, 11]. High cognitive load (e.g., time pressure in the clinical decision-making context or tasks in the financial context), can impair the user's ability to respond to a complex explanation and lead to lower EUI, even for accurate or relevant explanations. When it comes to high-load settings, explanation designs should focus on short explanations that deliver the most critical evidence, rather than an exhaustive attribution.

A major risk in human-AI collaborations is automation bias, whereby humans uncritically accept the action proposal. By providing a verifiable reasoning trace, XAI has the potential to avoid this effect, as it can help human users adopt an engaged rather than a passive decision-making role. Explanation designs that require the user to evaluate an AI's explanation before accepting its recommendation, called explanation-gated acceptance, have been shown to reduce automation bias without considerably increasing decision time. Salimparsa et al. [20] reviewed XAI in clinical decision support systems and identified explanation-gated acceptance as a high-priority design direction for future research.

6. Empirical Analysis

6.1 Scenario Configuration

To demonstrate the application of the TCS, EUI, and HACG measures, this section presents illustrative numerical analyses across three conditions and three application scenarios. The three conditions represent different XAI deployment states: XAI-enabled with SHAP explanations targeting domain expert users; XAI-enabled with LIME explanations targeting non-expert users; and no XAI, representing a black-box model deployment. Input parameters are calibrated to empirical ranges reported in the studies referenced throughout this article, including trust and accuracy figures from Wang et al. [11], Naiseh et al. [10], and Agrawal et al. [18].

Methodological note. The analyses in this section are illustrative rather than experimentally derived. The parameter values are drawn from published empirical ranges and are used to validate formula behavior and demonstrate diagnostic application across realistic operating conditions. This approach is appropriate for introducing a novel evaluation framework, as it allows systematic exploration of the metrics' behavior across controlled conditions. However, the absence of primary experimental data is a limitation; the analyses should be interpreted as demonstrations of framework application rather than empirical findings. Section 8 identifies controlled user studies as the highest priority for future validation.

6.2 TCS and EUI Analysis

Table 2 presents TCS and EUI values for the three conditions. In Condition 1, with SHAP explanations and domain expert users, user trust $T_u = 0.83$ closely tracks actual system accuracy $A_s = 0.85$:

$$TCS = 1 - |0.83 - 0.85| = 1 - 0.02 = 0.980$$

EUI is computed with comprehension gain $\Delta C = 0.79$ and decision quality improvement $\Delta D = 0.66$, with domain-expert weights $w^c = 0.60$ and $w^d = 0.40$:

$$EUI = (0.79 \times 0.60 + 0.66 \times 0.40) / 1.0 = (0.474 + 0.264) / 1.0 = 0.738$$

In Condition 2, with LIME explanations and non-expert users, trust and accuracy values are lower and more closely aligned: $T_u = 0.74$, $A_s = 0.72$, $TCS = 0.980$. The EUI reflects the reduced explanation effectiveness for non-expert users: $\Delta C = 0.45$, $\Delta D = 0.38$, $EUI = 0.422$.

In Condition 3, with no XAI, users exhibit over-trust: $T_u = 0.91$ against $A_s = 0.65$, yielding:

$$TCS = 1 - |0.91 - 0.65| = 1 - 0.26 = 0.740$$

With no explanation available, comprehension gain and decision improvement are minimal: $\Delta C = 0.18$, $\Delta D = 0.12$, producing $EUI = 0.156$, falling in the low utility zone. Table 2 consolidates all values.

Table 2. TCS and EUI Results Across Three XAI Deployment Conditions

Condition	T_u	A_s	TCS	EUI	Classification
XAI-enabled, SHAP (domain experts)	0.83	0.85	0.980	0.738	Well-calibrated trust; high explanation utility
XAI-enabled, LIME (non-expert users)	0.74	0.72	0.980	0.422	Well-calibrated trust; moderate explanation utility
No XAI (black-box model)	0.91	0.65	0.740	0.156	Moderate miscalibration (over-trust); low explanation utility

The results confirm a consistent pattern: XAI-enabled conditions produce substantially better trust calibration and explanation utility than black-box conditions. The SHAP condition produces a 32.4% improvement in TCS over the no-XAI condition, and a 373% improvement in EUI, confirming that explanation quality has a disproportionate effect on utility relative to trust calibration. This asymmetry suggests that in practice, trust

calibration improvements are achievable through relatively simple transparency measures, while explanation utility gains require more substantial investment in explanation design and user interface.

6.3 HACG-Based Collaboration Assessment

Table 3 presents HACG calculations for three application scenarios. In the clinical diagnosis scenario with SHAP explanations, human performance $P_{\text{human}} = 0.74$, AI performance $P_{\text{AI}} = 0.84$, and collaborative performance $P_{\text{collab}} = 0.91$:

$$\text{HACG} = (0.91 - \max(0.74, 0.84)) / \max(0.74, 0.84) = (0.91 - 0.84) / 0.84 = 0.083$$

A positive HACG of 0.083 indicates that the XAI-enabled team outperforms the AI-only baseline by 8.3%. In the fraud detection scenario with LIME:

$$\text{HACG} = (0.87 - 0.81) / 0.81 = 0.074$$

The collaborative team achieves a 7.4% gain over the AI-only baseline, a positive but smaller gain reflecting the lower explanation utility of LIME for non-expert users in this domain. In the document review scenario with no XAI:

$$\text{HACG} = (0.79 - 0.82) / 0.82 = -0.037$$

The collaboration deficit of 3.7% arises because users in the no-XAI condition exhibit automation bias: they accept AI recommendations without critical assessment, importing the AI's errors into the combined decision stream while contributing little corrective value. This negative HACG confirms the theoretical prediction that XAI is a necessary condition for productive human-AI collaboration in high-stakes decision domains.

Table 3. HACG Results Across Three Application Scenarios

Scenario	P_human	P_AI	P_collab	HACG	Classification
Clinical diagnosis (SHAP enabled)	0.74	0.84	0.91	+0.083	Complementary gain (8.3% above AI baseline)
Fraud detection (LIME enabled)	0.69	0.81	0.87	+0.074	Complementary gain (7.4% above AI baseline)
Document review (no XAI)	0.76	0.82	0.79	-0.037	Collaboration deficit (3.7% below AI baseline)

7. Applications

XAI is applied across a variety of areas; in clinical decision support systems, XAI allows clinicians to verify the diagnosis generated by the AI with their own clinical assessment. Stiglic et al. [15] observed that using ML explanation methods increased calibration of clinician trust and confidence for decision-making in an intensive care and readmission prediction setting using SHAP-based explanations. In a follow-up study by Agrawal et al. [18], adding XAI methods for a disease prediction model in clinical AI applications improved decision transparency and regulatory compliance.

The XAI module can also address regulatory concerns regarding explainable credit and fraud predictions. SHAP-based feature attributions can be useful to loan officers or risk analysts for both interpreting a specific credit decision and monitoring model performance at the aggregate level. Counterfactual explanations [13] are also particularly useful, as they can show the applicant what would have to change in order for a different outcome to take place. This makes them compatible with the GDPR Article 22 requirement for meaningful explanation of automated decision-making, according to Salih et al. [14], who found SHAP and LIME to be complementary. SHAP was recommended for regulatory reporting, and LIME for real-time user interaction. Recommender systems are another common domain. Because explanations of recommendations directly affect users' adherence to and acceptance of these recommendations, they comprise one of the top-volume use cases for XAI. Naiseh et al. [10] present evidence that explanation-improved recommender systems lead to more appropriate user reliance (including less over-acceptance of incorrect recommendations and under-acceptance of correct recommendations). Wang et al. [11] showed that users who receive explanations for AI recommendations make more accurate predictions of recommendation quality, enabling better-calibrated adoption decisions. Table 4 summarizes these and other application domains with their primary XAI methods and trust dimensions.

Table 4. XAI Application Domain Summary with Primary Methods and Trust Dimensions

Application Domain	Primary XAI Method(s)	Key Trust Dimension	Representative Reference
Clinical decision support	SHAP, LIME, counterfactual	Diagnostic accuracy trust	Stiglic et al. [15]; Agrawal et al. [18]
Financial risk assessment	SHAP, interpretable models	Risk calibration trust	Wachter et al. [13]; Salih et al. [14]

Recommender systems	LIME, feature importance	Recommendation acceptance trust	Naiseh et al. [10]; Wang et al. [11]
Autonomous systems	Grad-CAM, saliency maps	Safety and reliability trust	Adadi and Berrada [7]; Gunning and Aha [9]

Autonomous systems, including autonomous vehicles, robotic process automation, and AI-assisted inspection systems, represent an application domain where trust miscalibration carries direct safety consequences. XAI methods designed for this domain, including Grad-CAM [12] for visual perception systems and rule extraction for control system models, enable human supervisors to understand when and why the system is operating outside its reliable performance envelope. The DARPA XAI program [9] prioritized this application domain precisely because safety-critical human oversight requires sufficient transparency to support reliable trust calibration.

8. Limitations and Future Directions

Despite the analytical and practical contributions of this article, the proposed framework carries limitations that must be acknowledged. The three measures, TCS, EUI, and HACG, rely on operationalizations of trust, comprehension, and performance that are application-specific and may not generalize directly across domains. In particular, the user trust score T_u in TCS may be difficult to measure reliably in operational settings: self-report trust scales are subject to social desirability bias, while behavioral proxies such as recommendation acceptance rates conflate trust with convenience, task urgency, and decision habit. Practitioners implementing TCS in real-world deployments should validate the trust measurement instrument against behavioral benchmarks.

The accuracy-interpretability trade-off represents a persistent structural challenge in XAI research [4, 5]. Many high-accuracy models, including deep neural networks and gradient-boosted ensembles, are difficult to explain precisely because their decision surfaces are complex and high-dimensional. Post-hoc explanation methods such as LIME and SHAP provide approximate local accounts of these surfaces but may produce explanations that are accurate on average while being misleading for specific instances. The risk of misleading explanations is particularly acute in high-stakes decisions, where a user who acts on an inaccurate explanation may make a worse decision than one who ignores the model entirely. Salih et al. [14] identified this phenomenon as a primary limitation of current post-hoc explanation methods.

The illustrative analyses in Section 6 use values calibrated to published benchmarks rather than primary experimental data. This represents the most significant limitation of the current work. While the illustrative approach is appropriate for introducing and validating a novel framework's mathematical properties, empirical validation through controlled user studies is the essential next step. Such studies should systematically vary explanation modality, user expertise, and application domain to confirm the diagnostic patterns observed in the illustrative analyses and to establish normative ranges for TCS, EUI, and HACG under realistic operating conditions.

Beyond empirical validation, the most productive future research directions, in order of priority, are: (1) controlled user studies across multiple domains to validate TCS-EUI-HACG diagnostic patterns; (2) adaptive explanation systems that dynamically select explanation modality and complexity based on real-time user cognitive load assessment; (3) integration with self-supervised evaluation methods that measure explanation quality without requiring human annotation; and (4) extension of HACG to multi-agent human-AI teams with multiple humans and multiple AI systems [19]. The framework also raises open questions about measuring warranted trust over extended operational periods where AI performance may drift [17].

9. Conclusion

This article introduces a formal evaluation framework for explainable artificial intelligence in human-AI collaborative decision-making systems. The three measures, Trust Calibration Score, Explanation Utility Index, and Human-AI Collaboration Gain, provide a compact and operationally grounded set of diagnostics for assessing XAI effectiveness across the dimensions most critical to human-AI collaboration: trust alignment, explanation quality, and collaborative performance.

Illustrative numerical analyses showed that XAI-enabled conditions lead to much better trust calibration and explanation utility than black-box deployments, with SHAP-based explanations for domain expert users achieving $TCS = 0.980$ and $EUI = 0.738$ compared to $TCS = 0.740$ and $EUI = 0.156$ in the no-XAI condition. The HACG analyses confirmed that effective XAI is a necessary condition for productive human-AI collaboration: XAI-enabled teams achieved collaboration gains of 7 to 9 percent above the AI-only baseline, while the no-XAI team exhibited a collaboration deficit of 3.7 percent due to automation bias and miscalibrated trust.

These findings reinforce a central argument of the XAI field: technical accuracy is necessary but not sufficient for beneficial AI deployment. Decision systems must also be designed for human understanding, with explanation mechanisms calibrated to the expertise, cognitive capacity, and decision context of their intended users. The TCS-EUI-HACG framework offers researchers and practitioners a replicable methodology for pursuing this design goal, and its further development represents a productive direction for advancing responsible human-AI collaboration.

References

- [1] Finale Doshi-Velez and Been Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017. Available: <https://arxiv.org/abs/1702.08608>
- [2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "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, 2016. Available: <https://dl.acm.org/doi/abs/10.1145/2939672.2939778>
- [3] Scott M. Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017. Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [4] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl, "Interpretable machine learning: a brief history, state-of-the-art and challenges," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 417-431, Cham: Springer International Publishing, 2020. Available: https://link.springer.com/chapter/10.1007/978-3-030-65965-3_28
- [5] Zachary C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31-57, 2018. Available: <https://spawn-queue.acm.org/doi/pdf/10.1145/3236386.3241340>
- [6] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071-22080, 2019. Available: <https://www.pnas.org/doi/abs/10.1073/pnas.1900654116>
- [7] Amina Adadi and Mohammed Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138-52160, 2018. Available: <https://ieeexplore.ieee.org/abstract/document/8466590>
- [8] Tim Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2019. Available: <https://www.sciencedirect.com/science/article/pii/S0004370218305988>
- [9] David Gunning and David W. Aha, "DARPA's explainable artificial intelligence program," *AI Magazine*, vol. 40, no. 2, pp. 44-58, 2019. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1609/aimag.v40i2.2850>
- [10] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali, "Explainable recommendation: when design meets trust calibration," *World Wide Web*, vol. 24, no. 5, pp. 1857-1884, 2021. Available: <https://link.springer.com/article/10.1007/s11280-021-00916-0>
- [11] Xinru Wang, Zhuoran Lu, and Ming Yin, "Will you accept the AI recommendation? Predicting human behavior in AI-assisted decision making," in *Proceedings of the ACM Web Conference 2022*, pp. 1697-1708, 2022. Available: <https://dl.acm.org/doi/abs/10.1145/3485447.3512240>
- [12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618-626, 2017. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html
- [13] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law and Technology*, vol. 31, p. 841, 2017. Available: <https://doi.org/10.2139/ssrn.3063289>
- [14] Ahmed M. Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Karim Lekadir, and Gloria Menegaz, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, 2025. Available: <https://advanced.onlinelibrary.wiley.com/doi/full/10.1002/aisy.202400304>
- [15] Gregor Stiglic, Primož Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Wiley Interdisciplinary*

- Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 5, p. e1379, 2020. Available: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1379>
- [16] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan, "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, 2020. Available: <https://dl.acm.org/doi/abs/10.1145/3313831.3376219>
- [17] Regina de Brito Duarte, Filipa Correia, Patricia Arriaga, and Ana Paiva, "AI trust: Can explainable AI enhance warranted trust?" Human Behavior and Emerging Technologies, vol. 2023, no. 1, p. 4637678, 2023. Available: <https://onlinelibrary.wiley.com/doi/full/10.1155/2023/4637678>
- [18] Renuka Agrawal, Tawishi Gupta, Shaurya Gupta, Sakshi Chauhan, Prisha Patel, and Safa Hamdare, "Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency," Diagnostic Pathology, vol. 20, no. 1, p. 105, 2025. Available: <https://link.springer.com/article/10.1186/s13000-025-01686-3>
- [19] Gerui Xu, Shruthi Venkatesha Murthy, and Bochen Jia, "Enhancing intuitive decision-making and reliance through human-AI collaboration: A review," Informatics, vol. 12, no. 4, p. 135, 2025. Available: <https://www.mdpi.com/2227-9709/12/4/135>
- [20] Mozghan Salimparsa, Kamran Sedig, Daniel J. Lizotte, Sheikh S. Abdullah, Niaz Chalabianloo, and Flory T. Muanda, "Explainable AI for clinical decision support systems: literature review, key gaps, and research synthesis," Informatics, vol. 12, no. 4, p. 119, 2025. Available: <https://www.mdpi.com/2227-9709/12/4/119>