

# Building Trust in AI: Explainable AI in Critical Decision-Making

Ingle Sakshi Atmaram, Prof. Bangar Abhishek Popat, Prof. Bhosale Sachin Bajirao, Prof. Karad V. A.

Dept. of Computer Engineering  
Jaihind College of Engineering, Pune, India

**Abstract** — Artificial intelligence (AI) is increasingly used to support decisions in healthcare, finance, recruitment, transportation, and public services. Although complex machine-learning models can produce accurate predictions, their reasoning is not always easy to understand. This creates practical questions: why did the system recommend a particular outcome, did it use appropriate information, and when should a person challenge its decision? Explainable artificial intelligence (XAI) brings together techniques that help people understand a model's behaviour or the factors behind a prediction. This review discusses interpretable models, feature-attribution methods such as LIME and SHAP, counterfactual explanations, and model-agnostic approaches. It compares their uses and limitations in high-impact settings. An understandable explanation, however, is not proof that a model is accurate, fair, dependable, or safe. The paper uses a narrative review of foundational work and selected research and survey publications on interpretability, responsible AI, and human-centred explanations. No new experiments, participant studies, or quantitative tests were performed. The literature points to unresolved issues involving explanation fidelity, stability, accessibility, privacy, bias, evaluation standards, and over-reliance on automated advice. Future research should use domain-specific evaluations, retain meaningful human oversight, communicate uncertainty clearly, assess fairness across groups, and report limitations transparently. XAI can support accountability and informed supervision, but it cannot guarantee trustworthiness on its own.

**Keywords**— Explainable Artificial Intelligence, XAI, Critical Decision-Making, Interpretability, Trustworthy AI, Human-AI Interaction, Accountability, Model Transparency.

## I. INTRODUCTION

AI increasingly influences decisions about health, financial security, employment, travel, and access to public services. Deep neural networks and other complex models can identify patterns in large datasets, but their internal decision processes may be difficult to inspect. In low-risk situations, this can be inconvenient; in high-impact settings, it can make accountability, appeals, and safe use more difficult. Explainable artificial intelligence (XAI) aims to provide useful information about how an AI system works or why it produced a particular output. Interpretability and explainability are related but distinct: an interpretable model is structured so that people can understand it directly, while an explanation method may be applied after a complex model has been trained. Common approaches include feature importance, local explanations, counterfactual examples, and models designed to be understandable [1], [5]. The objective should not be to increase trust in every AI system. Instead, trust should be calibrated to the evidence: people should rely on a system when its performance and limitations justify doing so, and question its advice when concerns arise. An explanation may sound persuasive while remaining incomplete or misleading. XAI must therefore be considered alongside accuracy, robustness,

fairness, privacy, and effective human supervision. This paper reviews established XAI methods for consequential decisions, compares their practical trade-offs, identifies unresolved challenges, and outlines research directions for safer cooperation between people and AI.

This paper reviews established XAI methods and their relevance to high-impact decisions. It summarizes key techniques, compares their practical trade-offs, highlights unresolved issues, and identifies research directions that may support safer cooperation between people and AI systems.

## II. REVIEW METHODOLOGY AND LITERATURE SELECTION CRITERIA

This paper uses a narrative literature-review approach rather than a systematic review or meta-analysis. The selected sources cover foundational explanation methods, surveys of XAI, interpretability, counterfactual explanations, and research into the relationship between explanations and human trust. Priority was given to peer-reviewed journal and conference papers, along with a widely used reference book on interpretable machine learning.

Sources were considered relevant when they addressed at least one of these themes: (i) definitions or classifications of explainability and interpretability; (ii) local or global explanation methods; (iii) counterfactual explanations and responses to automated decisions; (iv) the relationship between explanations, trust, and human decision-making; or (v) ethical, legal, and technical limits of XAI. The purpose is to provide a thematic overview, not an exhaustive catalogue of every publication. No formal database-search protocol, duplicate-screening process, or numerical study-quality assessment was applied. The paper therefore does not claim the completeness expected of a systematic review.

The discussion draws on Ribeiro et al.'s work on LIME [1], Lundberg and Lee's SHAP method [2], Lipton's analysis of interpretability [3], and surveys by Adadi and Berrada [4] and Arrieta et al. [5]. Wachter et al. [6] inform the section on counterfactual explanations in automated decisions. Wiratsin and Ragkhitwetsagul [7] are considered for their review of XAI effectiveness and human trust, while Molnar's online book [8] provides additional methodological background.

### III. LITERATURE REVIEW

#### 1. Interpretable Models and the Meaning of Explanation

Interpretability has no single definition that is accepted across all fields, nor is there one universal way to measure it. Lipton [3] explains that the term can refer to several aims, including transparency, trust, and scientific understanding. Small decision trees, linear models, and rule-based systems are often easier to inspect directly. However, they are useful only when their assumptions suit the task and their predictive performance is adequate.

#### 2. Local Explanation Methods: LIME

LIME (Local Interpretable Model-agnostic Explanations) explains an individual prediction by fitting a simpler, interpretable model around the case being examined [1]. It perturbs the input, observes the complex model's outputs, and builds a local approximation of its behaviour. Because LIME is model-agnostic, it can be applied to different model types. Its explanations can nevertheless vary with the perturbation strategy, feature representation, and choice of local neighbourhood. A local explanation should not be mistaken for a complete description of the model's behaviour in all situations.

#### 3. Feature Attribution: SHAP

SHAP (SHapley Additive exPlanations) uses concepts from cooperative game theory to estimate how much each feature

contributes to a prediction [2]. Its additive format provides a consistent way to display feature contributions for individual cases and broader summaries. In practice, computation costs, relationships among features, baseline choices, and the interpretation of attribution values can complicate its use. An attributed contribution is not the same as a causal effect; on its own, it does not prove what would happen if a feature were changed.

#### 4. Counterfactual Explanations

Counterfactual explanations show how an input would have to change for a model to produce a different outcome. For instance, an automated decision system might indicate that an application would receive another result if one or more eligible values were different. Wachter et al. [6] discuss this approach in the context of automated decisions and legal considerations. Counterfactuals may help people understand possible routes to a different result, but suggested changes must be realistic and appropriate. They may be impossible to achieve, depend on an unchangeable attribute, or fail to reflect the true causal process.

#### 5. Surveys and Responsible AI

Adadi and Berrada [4] survey the growing XAI field and explain why researchers seek ways to inspect black-box models. Arrieta et al. [5] place explainability within broader responsible-AI goals, including transparency and accountability. Together, these surveys show that XAI is not one universal technique. The most suitable approach depends on the model, audience, task, and consequences of an incorrect decision. Explanations should therefore be designed and evaluated for the setting in which they will be used.

#### 6. Explanations and Human Trust

Explanations are often expected to increase trust, but the connection is not simple. Wiratsin and Ragkhitwetsagul [7] review evidence on how well XAI techniques improve people's trust in machine-learning models. In high-stakes applications, perceived persuasiveness must be separated from justified trust. A clear explanation may make users more confident even when it is unstable or the prediction is wrong. Evaluations should test whether explanations help people spot errors, understand limitations, make better decisions, and recognize when AI advice should be accepted or challenged.

### IV. COMPARATIVE ANALYSIS OF EXISTING STUDIES

Table 1 summarizes selected publications and their contributions to XAI. The works are not directly interchangeable: some explain individual predictions, some

clarify key concepts or survey the field, and others examine how explanations relate to trust.

Table 1. Comparison of selected XAI publications

| Study                                 | Approach / contribution                                        | Limitation / consideration                                                   |
|---------------------------------------|----------------------------------------------------------------|------------------------------------------------------------------------------|
| Ribeiro et al. (2016) [1]             | LIME; local surrogate explanations for individual predictions. | A local approximation may be unstable and does not describe the whole model. |
| Lundberg & Lee (2017) [2]             | SHAP; additive feature-attribution framework.                  | Attributions depend on assumptions and do not establish causation.           |
| Lipton (2018) [3]                     | Conceptual analysis of interpretability.                       | There is no universally accepted definition or evaluation measure.           |
| Adadi & Berrada (2018) [4]            | Survey of XAI motivations, methods, and challenges.            | Does not identify one method that is best for every domain.                  |
| Arrieta et al. (2020) [5]             | Connects XAI concepts and taxonomies with responsible AI.      | An explanation alone cannot ensure responsible deployment.                   |
| Wachter et al. (2018) [6]             | Counterfactual explanations for automated decisions.           | Suggested changes may be infeasible or non-causal.                           |
| Wiratsin & Ragkhitwetsagul (2025) [7] | Systematic review of XAI methods and human trust.              | Effects may differ across tasks, users, and evaluation measures.             |

## V. RESEARCH CHALLENGES AND IDENTIFIED GAPS

### 1. Faithfulness and Explanation Quality

An explanation should accurately represent the behaviour of the model it claims to describe. A simple substitute may be easy to understand yet fail to capture the original model adequately. Conversely, a technically faithful explanation may be too complex for its intended audience. Evaluation should balance fidelity, simplicity, completeness, and usability; no single measure is likely to capture every important quality.

### 2. Stability and Reproducibility

Minor changes to an input, model, or sampling procedure can produce different explanations. If an explanation changes

substantially while the prediction remains similar, users may struggle to rely on it. Researchers should report stability, sensitivity to settings, and reproducibility across repeated runs.

### 3. Human Factors and Trust Calibration

Some evaluations measure whether people say they trust an AI system instead of checking whether they use it appropriately. A polished explanation can encourage excessive confidence in a wrong recommendation. Further studies should measure error detection, decision quality, confidence calibration, and whether users recognize when independent human review is needed.

### 4. Fairness, Privacy, and Security

Explanations can reveal sensitive attributes or private information, and may expose details that could be used to manipulate a model. They can also make biased patterns visible without correcting those patterns automatically. High-impact applications should pair XAI with fairness assessments, privacy protections, access controls, and testing for adversarial misuse.

### 5. Domain and Audience Dependence

Different audiences need different explanations. A clinician, loan officer, job applicant, regulator, and machine-learning engineer may each need a different level of detail. One generic interface is unlikely to meet every need. The application domain and feedback from intended users should shape both the content and presentation of explanations.

### 6. Lack of Shared Evaluation Standards

Studies use different datasets, participants, explanation methods, and measures of trust or usefulness, which makes their findings difficult to compare. Better benchmarks should include realistic decision tasks, diverse users, appropriate baselines, and clear reporting of failures as well as successes.

## VI. PROPOSED RESEARCH DIRECTIONS

The literature reviewed suggests several directions that may improve the practical value of XAI in critical decision-making. These are proposals for future research, not results from experiments conducted for this paper.

- Measure appropriate reliance, not trust alone. Examine whether people accept, question, or override AI recommendations appropriately, including whether they can identify deliberately introduced errors.
- Show uncertainty alongside explanations. Communicate prediction confidence or uncertainty in accessible language and test whether this helps users recognize

ambiguous cases or inputs that differ from the training data.

- Evaluate outcomes within each domain. In healthcare, finance, recruitment, and other fields, test whether explanations improve decision quality, fairness, and opportunities to challenge a decision—not merely user satisfaction.
- Test explanation stability and fidelity. Report how explanations respond to changes in inputs, background data, model versions, and explanation settings, and clearly describe the limits of the evaluation measures.
- Keep human review in high-risk workflows. Define when an explanation should trigger further checks, escalation, or an independent human decision.
- Make counterfactual suggestions realistic and fair. Recommend changes only to feasible and legitimate features, and assess whether the suggested actions place unequal burdens on different groups.
- Report methods and limitations transparently. Document the model's intended use, data limitations, explanation method, evaluation conditions, known failure modes, and situations requiring human oversight.

## VII. CONCLUSION AND FUTURE SCOPE

Explainable AI can help people examine AI-supported decisions, understand relevant factors, and challenge outcomes when appropriate. LIME, SHAP, and counterfactual explanations offer different kinds of insight, while broader surveys connect XAI with responsible AI. However, an explanation is not complete transparency, and a convincing explanation does not establish that a model is accurate, fair, robust, or safe. Important concerns remain around fidelity, stability, privacy, fairness, audience needs, and evaluation.

Future work should use realistic, domain-specific studies to determine whether explanations improve decisions and help people rely on AI appropriately. Other priorities include communicating uncertainty, strengthening human oversight, enabling practical recourse, and developing shared reporting standards.

Because this paper is a narrative review and includes no original experiments, its conclusions synthesize selected literature rather than presenting new empirical evidence. Trust in critical AI systems must ultimately be earned through demonstrated performance, accountability, and effective oversight. XAI can support these goals, but cannot replace them.

## REFERENCES

1. M. T. Ribeiro, S. Singh, and C. 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, 2016, pp. 1135–1144.
2. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
3. Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.
4. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.
5. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
6. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.
7. I.-O. Wiratsin and C. Ragkhitwetsagul, "Effectiveness of Explainable Artificial Intelligence (XAI) Techniques for Improving Human Trust in Machine Learning Models: A Systematic Literature Review," IEEE Access, vol. 13, pp. 121326–121350, 2025.
8. C. Molnar, Interpretable Machine Learning, 2nd ed., 2022. Available: <https://christophm.github.io/interpretable-ml-book/>.