American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

Explainable AI in High-Stakes Decision Making: Building Transparency, Accountability and Public Trust

Eric von Hippel

Professor Massachusetts Institute of Technology

Abstract

Artificial Intelligence (AI) is increasingly being deployed in high-stakes environments such as healthcare, financial services, criminal justice, employment, insurance, public administration, and critical infrastructure. While AI systems can improve prediction, efficiency, and decision consistency, their use in consequential settings raises concerns regarding opacity, bias, accountability, privacy, and public trust. Explainable Artificial Intelligence (XAI) has emerged as an important approach for addressing these concerns by providing information about how AI systems generate predictions or recommendations. This study examines the role of explainable AI in high-stakes decision making, with particular emphasis on transparency, accountability, and public trust. A qualitative and conceptual methodology based on secondary literature is adopted to analyse different approaches to explainability, including interpretable models, feature attribution, counterfactual explanations, example-based explanations, and human-in-the-loop oversight. The study proposes an integrated framework in which technical explainability is combined with procedural accountability, institutional governance, human oversight, and stakeholder communication. The analysis suggests that explanations can improve understanding and contestability, but explainability alone cannot guarantee fairness or trustworthy AI. Excessively complex explanations may also create false confidence or fail to address underlying biases. Effective governance therefore requires a broader approach incorporating documentation, auditing, impact assessment, monitoring, appeal mechanisms, data governance, and clearly assigned responsibility. The study concludes that explainable AI should be understood not merely as a technical feature but as a component of responsible decision-making architecture. In high-stakes environments, meaningful transparency must enable affected stakeholders to understand, question, and appropriately challenge AI-supported decisions.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

Keywords: Explainable Artificial Intelligence, High-Stakes Decision Making, Transparency, Accountability, Public Trust, AI Governance, Algorithmic Fairness, Human Oversight, Responsible AI, Artificial Intelligence.

1. Introduction

Artificial Intelligence has moved from experimental applications into decision-making systems that can directly affect people's lives.

AI is increasingly used to support decisions involving:

- Medical diagnosis.
- Credit approval.
- Insurance assessment.
- Employee recruitment.
- University admissions.
- Criminal justice.
- Welfare eligibility.
- Fraud detection.
- Immigration assessment.
- Public-sector resource allocation.

These applications differ substantially from low-risk AI applications such as entertainment recommendations or image enhancement.

When an AI decision can affect a person's health, freedom, employment, financial security, or access to public services, trust and accountability become essential.

A central challenge is that many advanced AI systems, particularly complex machine-learning models, can be difficult for humans to understand.

This creates a fundamental problem:

AI Decision → Outcome

but stakeholders may not know:

Why did the system reach this outcome?

Explainable AI seeks to address this problem by providing meaningful information about the factors, patterns, or reasoning associated with an AI output.

However, explainability is not simply about opening the "black box".

An effective explanation should enable stakeholders to:

- Understand relevant factors.
- Identify potential errors.
- Detect unfair treatment.
- Challenge inappropriate decisions.
- Assign responsibility.
- Make informed decisions.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

Therefore, XAI has become closely connected with the broader goals of responsible and trustworthy AI.

2. Concept of Explainable Artificial Intelligence

Explainable Artificial Intelligence refers broadly to methods and systems that make AI outputs understandable to relevant users.

Explainability can involve different levels.

Global Explainability

Explains how a model generally behaves.

Local Explainability

Explains why a particular prediction was generated.

Feature-Based Explanation

Identifies factors that contributed to an output.

Counterfactual Explanation

Explains what would need to change for a different outcome.

Example-Based Explanation

Uses similar cases to explain a prediction.

Different stakeholders may require different explanations.

For example, a data scientist may need technical model information, while a patient may simply need an understandable explanation of why an AI-supported recommendation was generated.

3. High-Stakes Decision Making

High-stakes decisions are decisions in which errors or unfair outcomes can produce significant consequences.

Examples include:

Healthcare

Incorrect diagnosis or treatment recommendation.

Finance

Unfair credit rejection.

Employment

Discriminatory recruitment decisions.

Criminal Justice
Incorrect risk assessment.

Public Administration
Improper denial of benefits.

Insurance
Unjustified premiums or claims decisions.
In these contexts, AI systems should generally operate with stronger governance and oversight than low-risk applications.

4. Transparency, Explainability and Accountability

These concepts are related but not identical.

Transparency
Concerns whether relevant information about the AI system is available.

Explainability
Concerns whether the system's outputs can be meaningfully understood.

Accountability
Concerns who is responsible for the system and its consequences.

Trust
Concerns whether stakeholders are willing to rely on the system appropriately.
A useful relationship is:

Transparency → Understanding → Contestability → Accountability → Trust
However, trust should not be created through explanations alone.
Trustworthy AI requires evidence that the system is accurate, fair, secure, and appropriately governed.

5. Research Objectives

The study aims to:

1. Examine the role of XAI in high-stakes decision making.
2. Analyse the relationship between explainability and transparency.
3. Examine how XAI can strengthen accountability.
4. Investigate the role of explanations in building public trust.
5. Identify major XAI techniques.
6. Examine limitations and risks associated with explainability.

7. Develop an integrated framework for trustworthy AI-supported decision making.

6. Research Methodology

This study adopts a qualitative and conceptual research methodology based on secondary literature.

Research concerning explainable AI, algorithmic fairness, AI governance, transparency, accountability, human–AI interaction, and public trust is conceptually analysed.

The study focuses on five major dimensions:

1. Technical explainability.
2. Institutional transparency.
3. Human oversight.
4. Accountability mechanisms.
5. Stakeholder trust.

These dimensions are integrated into a conceptual framework for high-stakes AI governance.

7. Why Explainability Matters in High-Stakes AI

Explainability is important because affected individuals may have legitimate questions about AI-supported decisions.

For example:

Why was my loan application rejected?

Why was this patient classified as high-risk?

Why was I not selected for this job?

Why was my eligibility for a public benefit denied?

If organisations cannot provide meaningful answers, affected individuals may perceive the system as arbitrary or unfair.

Explainability can therefore support procedural fairness by allowing decisions to be questioned and reviewed.

8. Major Explainable AI Techniques

8.1 Interpretable Models

Some models are inherently easier to understand.

Examples include:

- Decision trees.
- Linear regression.
- Rule-based systems.
- Generalised additive models.

In high-stakes settings, interpretable models may sometimes be preferable when they provide sufficient predictive performance.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

8.2 Feature Attribution

Feature-attribution methods estimate the contribution of individual features to a prediction.
Examples include:

- SHAP.
- LIME.
- Integrated Gradients.

These methods can help identify which variables influenced a particular output.

8.3 Counterfactual Explanations

Counterfactual explanations describe how an outcome could change.
For example:
"The application would have received a different assessment if the relevant financial condition had been different."
This can be particularly useful because it provides actionable information.

8.4 Example-Based Explanations

The system explains a decision by identifying similar cases.
For example:
"This case received a similar classification to other cases with comparable characteristics."

8.5 Rule-Based Explanations

Complex model outputs can sometimes be translated into simplified decision rules.
These explanations can be useful for non-technical users.

9. Table

1. Explainability Techniques

Technique	Primary Purpose	Potential Application
Decision trees	Direct interpretability	Healthcare
Linear models	Understandable relationships	Finance
Feature attribution	Identify influential factors	Risk assessment
LIME	Local model explanation	Classification
SHAP	Feature contribution	Credit/healthcare
Counterfactuals	Explain alternative outcomes	Lending/employment
Example-based methods	Compare similar cases	Public administration
Rule-based explanations	Simplify decision logic	Compliance

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

10. Explainability in Healthcare

Healthcare represents one of the most important applications of XAI.

AI can support:

- Medical diagnosis.
- Risk prediction.
- Medical image analysis.
- Treatment recommendations.
- Patient monitoring.

Healthcare professionals need to understand AI recommendations because medical decisions involve substantial consequences.

For example, an AI system may identify a patient as high-risk.

A clinician may need to know:

- Which factors contributed to the assessment?
- How reliable is the prediction?
- Is the patient similar to the cases used for training?
- Could the result reflect data bias?

Explainability can help clinicians critically evaluate AI recommendations rather than accepting them automatically.

11. Explainability in Financial Decision Making

Financial institutions increasingly use algorithms for:

- Credit scoring.
- Fraud detection.
- Loan assessment.
- Risk management.
- Insurance decisions.

If an individual is denied credit, an understandable explanation can provide insight into the decision.

However, financial explanations must avoid simply providing generic statements.

A meaningful explanation should provide information that allows individuals and institutions to identify relevant factors and, where appropriate, correct inaccurate information.

12. Explainability in Employment

AI is increasingly used in recruitment and workforce analytics.

Applications include:

- CV screening.
- Candidate ranking.
- Skills assessment.
- Employee performance analysis.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

Because employment decisions affect livelihoods and career opportunities, explainability is important.

Organisations should be able to examine whether AI systems:

- Use relevant criteria.
- Produce consistent outcomes.
- Introduce discriminatory patterns.
- Exclude particular groups unfairly.

Human review remains important for consequential employment decisions.

13. Explainability in Criminal Justice

AI-supported risk assessments may be used in areas such as:

- Recidivism prediction.
- Bail assessment.
- Sentencing support.
- Resource allocation.

These applications require particularly strong safeguards.

An individual should not be disadvantaged by an opaque system without appropriate mechanisms for review and challenge.

Explainability can support:

- Judicial oversight.
- Legal scrutiny.
- Procedural fairness.
- Error detection.

14. Explainability in Public Administration

Governments may use AI to support:

- Welfare administration.
- Tax compliance.
- Fraud detection.
- Immigration processing.
- Public-resource allocation.

Public institutions have a special responsibility because citizens may have limited ability to opt out of government systems.

Consequently, AI-supported public decisions should generally include:

Transparency + Human Oversight + Appeal Mechanisms + Accountability

15. Human Oversight

Explainability should not replace human responsibility.

A human-in-the-loop model can be represented as:

AI Analysis → Explanation → Human Review → Decision → Monitoring

Human oversight is especially important when:

- The decision has high consequences.
- The model has uncertainty.
- Data quality is questionable.
- The decision affects vulnerable individuals.
- The individual contests the decision.

16. Accountability

Accountability requires clear responsibility throughout the AI lifecycle.

Responsibility may involve:

- Developers.
- Data scientists.
- Vendors.
- Managers.
- Deploying organisations.
- Decision-makers.

An organisation should be able to answer:

1. Who developed the system?
2. Who approved its deployment?
3. What data was used?
4. How was performance evaluated?
5. Who monitors the system?
6. Who reviews contested decisions?
7. Who is responsible when the system causes harm?

Explainability becomes meaningful only when it is connected to these accountability structures.

17. Public Trust

Public trust is influenced by perceptions of:

- Fairness.
- Transparency.
- Competence.
- Accountability.
- Privacy.
- Human control.

People are more likely to accept AI-supported decisions when they believe that the system is understandable, appropriately governed, and open to challenge.

However, excessive or poorly designed explanations may reduce trust.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

A technically accurate explanation that users cannot understand may have little practical value.

18. Explanation Quality

A good explanation should be:

Accurate

It should reflect the actual behaviour of the model.

Understandable

Non-technical stakeholders should be able to interpret it.

Relevant

It should focus on information relevant to the decision.

Actionable

Where appropriate, it should help users understand possible alternatives.

Timely

It should be available when decisions are being evaluated or challenged.

19. Table

2. Stakeholder-Specific Explanation Requirements

Stakeholder	Explanation Requirement
Data scientist	Model behaviour and technical diagnostics
Manager	Performance and risk
Doctor	Clinical factors and confidence
Customer	Decision factors and alternatives
Employee	Evaluation criteria
Regulator	Compliance and audit information
Judge	Evidence and decision logic
Citizen	Understandable reason for decision

20. Explainability and Algorithmic Fairness

Explainability and fairness are closely related but should not be treated as identical.

A model can be explainable but unfair.
For example, an understandable model may systematically disadvantage a particular group.
Therefore:

Explainability ≠ Fairness
A responsible AI system should combine:

- Explainability.
- Bias testing.
- Fairness assessment.
- Data-quality monitoring.
- Human oversight.

21. Explainability and Privacy
Greater transparency can sometimes conflict with privacy.
Revealing too much information about a model or dataset may expose:

- Personal information.
- Sensitive attributes.
- Proprietary algorithms.
- Security vulnerabilities.

Therefore, explanations should be designed to provide meaningful information without unnecessarily exposing confidential or personal data.

22. The Risk of False Transparency
One important challenge is the possibility of false transparency.
An organisation may provide an explanation that appears informative but does not actually allow meaningful understanding or contestability.
For example:
"The system rejected the application because the risk score was high."
This explains the outcome only superficially.
A more meaningful explanation would identify relevant decision factors and provide appropriate mechanisms for reviewing potentially incorrect information.

23. The Explainability–Accuracy Trade-Off
Some highly complex models may achieve strong predictive performance while being difficult to interpret.
Simpler models may be easier to understand but may not always provide equivalent predictive performance.
Therefore, organisations should not automatically assume:

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

More Explainable = Better
Instead, they should consider:

Accuracy + Explainability + Fairness + Risk + Context
The appropriate balance depends on the decision domain.

24. AI Auditing
AI auditing can evaluate whether a system operates as intended.
Audits may examine:

- Accuracy.
- Bias.
- Data quality.
- Model drift.
- Explainability.
- Security.
- Compliance.
- Human oversight.

Regular auditing is particularly important because model behaviour can change as data and environments evolve.

25. Proposed Conceptual Framework
The study proposes:

AI System
↓
Explainability Layer
↓
Human Understanding
↓
Human Review and Contestability
↓
Institutional Accountability
↓
Public Trust

The framework is supported by four governance mechanisms:

1. AI documentation.
2. Independent auditing.
3. Human oversight.
4. Appeal and redress mechanisms.

A broader representation is:

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

Data Governance

•

Model Governance

•

Explainability

•

Human Oversight

↓

Accountable AI Decision

↓

Fairness + Transparency + Contestability

↓

Public Trust

26. Data Analysis

The conceptual analysis identifies four major mechanisms through which XAI can contribute to trustworthy decision making.

1. Understanding

Stakeholders gain insight into decision factors.

2. Error Detection

Users can identify potentially incorrect or anomalous decisions.

3. Contestability

Affected individuals can challenge decisions.

4. Accountability

Institutions can establish responsibility for AI-supported outcomes.

These mechanisms collectively strengthen the legitimacy of AI deployment in high-stakes environments.

27. Challenges

27.1 Technical Complexity

Advanced models may be difficult to explain accurately.

27.2 Explanation Instability

Different explanation techniques may produce different interpretations.

27.3 User Understanding
Technical explanations may not be meaningful to ordinary users.

27.4 Model Bias
Explainability cannot automatically eliminate discriminatory patterns.

27.5 Privacy
Explanations may reveal sensitive information.

27.6 Intellectual Property
Organisations may be reluctant to disclose proprietary model information.

27.7 Over-Reliance
Users may trust AI recommendations too strongly when explanations appear convincing.

28. Recommendations
Organisations implementing AI in high-stakes environments should:

1. Conduct AI impact assessments before deployment.
2. Select explainability methods appropriate to the stakeholder.
3. Maintain human oversight over consequential decisions.
4. Conduct regular bias and performance audits.
5. Document model development and deployment.
6. Provide meaningful explanations to affected individuals.
7. Establish appeal and correction mechanisms.
8. Monitor model performance after deployment.
9. Protect personal and sensitive data.
10. Clearly assign organisational responsibility.
11. Train employees to interpret AI outputs critically.
12. Avoid presenting AI recommendations as automatically correct.

29. Questionnaire
5-Point Likert Scale: 1 = Strongly Disagree, 5 = Strongly Agree

No.	Statement
1	People should receive understandable explanations for high-stakes AI decisions.
2	Explainable AI increases confidence in AI-supported decisions.
3	AI systems used in high-stakes decisions should include human oversight.
4	Explainability can help identify potentially biased AI decisions.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

No.	Statement
5	Organisations should provide mechanisms for challenging AI-supported decisions.
6	Transparency is essential for responsible AI deployment.
7	AI decisions affecting individuals should be independently audited.
8	Technical explanations should be adapted to the needs of different stakeholders.
9	Explainability alone is insufficient to guarantee trustworthy AI.
10	Strong accountability mechanisms increase public acceptance of AI.

30. Future Research Directions

Future research should examine how different explanation formats affect public understanding and trust.

Potential areas include:

- XAI in healthcare.
- XAI in financial services.
- Explainable AI in public administration.
- Human–AI decision making.
- XAI and algorithmic fairness.
- XAI for vulnerable populations.
- Counterfactual explanations and decision contestability.
- AI auditing frameworks.
- Explainability and regulatory compliance.
- Public perceptions of AI transparency.

Experimental research could compare different explanation formats to determine which approaches most effectively improve understanding without generating excessive trust.

Longitudinal research could also investigate whether transparency produces sustained public trust or whether trust depends primarily on observed system performance.

31. Conclusion

Explainable Artificial Intelligence has become an important component of responsible AI governance, particularly when AI systems are used in decisions that can significantly affect individuals and communities.

In high-stakes environments, people need more than an automated decision. They need to understand why the decision was made, whether it is reasonable, who is responsible, and how it can be challenged.

XAI can contribute to these objectives through interpretable models, feature attribution, counterfactual explanations, example-based explanations, and human-in-the-loop systems.

However, explainability alone cannot guarantee trustworthy AI.
A system can be explainable and still be biased, inaccurate, insecure, or unfair.
Consequently, trustworthy high-stakes AI requires an integrated governance approach:
Explainability + Fairness + Transparency + Human Oversight + Auditing + Accountability + Redress
The ultimate objective should not be to make every AI system completely understandable to every person. Rather, organisations should provide meaningful, context-appropriate information that enables stakeholders to understand, evaluate, question, and appropriately challenge AI-supported decisions.
Public trust will ultimately depend not only on whether AI systems can explain their decisions, but on whether institutions demonstrate that those decisions are fair, accountable, contestable, and aligned with human values.

References

1. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
2. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
4. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
5. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.
6. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42.
7. Molnar, C. (2022). Interpretable Machine Learning. Independent publication.
8. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
9. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28, 689–707.
10. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399.
11. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

12. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.
13. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215.
14. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer.
15. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.