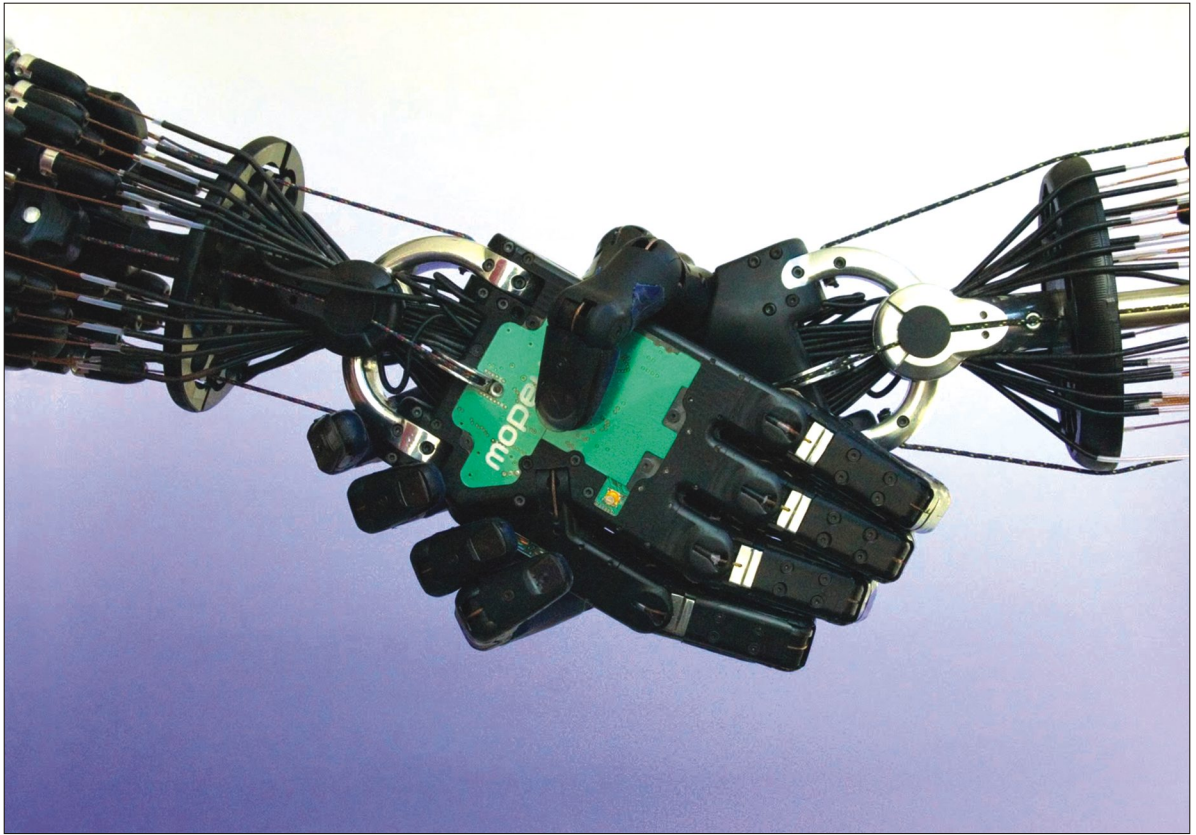




Journal of Artificial Intelligence

ISSN 1994-5450

science
alert

ANSI*net*
an open access publisher
<http://ansinet.com>

Systematic Review

Explainable Artificial Intelligence in High-Stakes Decision Systems: A Systematic Review (2015–2025)

Zeeshan Ashraf

Knowledge International Schools and Colleges, Faisalabad 38000, Pakistan

Abstract

Background and Objective: Artificial Intelligence systems are increasingly deployed in high-stakes domains, including healthcare, finance, and legal systems. The opacity of complex machine learning models raises concerns regarding accountability, fairness, and regulatory compliance. To systematically review Explainable Artificial Intelligence (XAI) methodologies, evaluation standards, and regulatory implications in high-risk decision systems from 2015 to 2025. **Materials and Methods:** A PRISMA-guided systematic review was conducted across Scopus, IEEE Xplore, ACM Digital Library, and Web of Science. Peer-reviewed journal articles published in English between January 2015 and January 2025 were included. **Results:** A total of 118 studies met the inclusion criteria. Healthcare accounted for 48% ($n = 57$) of studies, finance 32% ($n = 38$), legal systems 15% ($n = 18$), and other public-sector domains 5% ($n = 5$). A *post-hoc* explanation methods dominated (61%, $n = 72$), while intrinsic interpretable models comprised 39% ($n = 46$). Fidelity metrics were used in 74% ($n = 87$) of studies; however, only 27% ($n = 32$) conducted human-centered validation. **Conclusion:** Although, XAI research is expanding rapidly, standardized evaluation metrics and harmonized regulatory integration remain limited. Human-centered validation and domain-specific benchmarks are essential for responsible AI deployment in high-stakes systems.

Key words: Explainable AI, XAI, *post-hoc* methods, intrinsic models, high-stakes decision systems, healthcare AI, financial AI, regulatory compliance

Citation: Ashraf, Z., 2026. Explainable artificial intelligence in high-stakes decision systems: A systematic review (2015–2025). J. Artif. Intel., 19: 100-106.

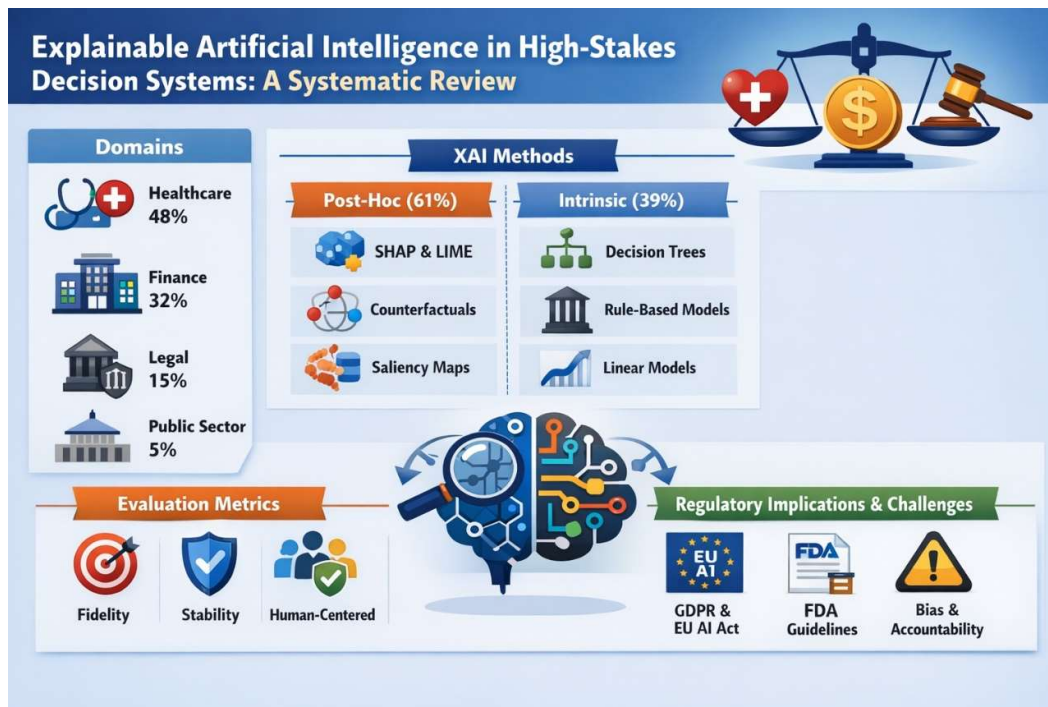
Corresponding Author: Zeeshan Ashraf, Knowledge International Schools and Colleges, Faisalabad 38000, Pakistan.

Copyright: © 2026 First Author *et al.* This is an open access article distributed under the terms of the creative commons attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

Graphical Abstract



INTRODUCTION

Artificial Intelligence (AI) systems are increasingly integrated into high-stakes decision-making across multiple domains, including healthcare diagnostics, credit scoring, fraud detection, and judicial risk assessment. The adoption of AI, particularly deep learning models, has substantially improved predictive performance, enabling more accurate and timely decisions. However, this performance often comes at the cost of interpretability, as complex “black-box” models obscure the reasoning behind their outputs. This opacity raises critical concerns regarding accountability, trust, and ethical responsibility, particularly in domains where decisions have significant human or societal consequences¹⁻⁴.

Explainability has thus emerged as a central research theme within AI, aimed at making model behavior transparent, comprehensible, and actionable for human stakeholders. Across disciplines, explainability encompasses a diverse set of approaches, ranging from feature attribution and saliency mapping to counterfactual reasoning and causal inference⁵⁻⁸. While some methods emphasize technical fidelity, others prioritize human-centered interpretability, reflecting the varying requirements of different domains and decision contexts^{3,7}. Despite this diversity, the conceptualization of

interpretability remains fragmented, with no universally accepted framework for measuring or operationalizing explanation quality.

In healthcare, Explainable AI (XAI) is particularly critical. Trust in clinical decision-support systems depends not only on predictive accuracy but also on the ability of clinicians to understand, verify, and justify recommendations in light of patient safety and ethical considerations^{9,10}. A *post-hoc* explanation techniques, such as SHAP, LIME, and saliency maps, are frequently employed to illuminate the predictions of deep learning models. Nevertheless, intrinsically interpretable models, such as decision trees and generalized additive models, are sometimes favored in life-critical settings to ensure transparent reasoning and facilitate regulatory compliance⁸.

In finance, AI-driven models are subject to stringent audit and compliance requirements, including anti-discrimination legislation and data protection regulations such as the General Data Protection Regulation (GDPR). Here, explainability ensures that automated decisions, such as credit scoring or fraud detection, are not only accurate but also legally defensible and fair¹¹. Similarly, in judicial systems, algorithmic opacity can undermine procedural fairness, challenge due process, and erode public trust. In these contexts, interpretable models are essential to prevent unintended biases, enhance transparency, and maintain legitimacy¹².

Despite these domain-specific needs, *post-hoc* interpretability approaches remain dominant across high-stakes applications. Their widespread use is largely driven by compatibility with complex deep learning architectures, which achieve superior predictive performance but are inherently opaque¹³⁻¹⁵. These approaches allow existing models to be augmented with explainability layers, providing insights into model behavior without requiring fundamental changes to the underlying architecture. However, reliance on *post-hoc* explanations introduces its own challenges, including potential instability, misleading attributions, and limited causal validity.

Three major challenges persist in XAI research, limiting its practical impact. First, there is a lack of standardized evaluation metrics, which hinders comparability across studies and complicates regulatory certification. Second, human-centered validation is limited, with relatively few studies assessing how explanations influence stakeholder decision-making, trust, or cognitive load. Third, regulatory alignment is insufficient: While legal and ethical frameworks increasingly mandate transparency and accountability, operationalizing these requirements remains technically ambiguous and inconsistently applied across domains.

This systematic review addresses these gaps by synthesizing peer-reviewed evidence from 2015 to 2025 on the application of XAI in high-stakes decision systems. The review examines methodological trends, domain-specific requirements, evaluation practices, and regulatory considerations, with the goal of providing a comprehensive overview of current capabilities and limitations. By highlighting best practices and identifying critical research needs, this study contributes to the development of more reliable, human-centered, and regulatory-aligned AI systems that can be safely deployed in environments where decisions carry substantial consequences.

MATERIALS AND METHODS

Study design: This systematic review was conducted following the PRISMA 2020 guidelines to ensure methodological transparency, reproducibility, and rigor. The approach was designed to synthesize evidence on Explainable AI (XAI) applications in high-stakes domains, emphasizing methodological diversity, human-centered evaluation, and regulatory alignment.

Data sources and search strategy: A comprehensive literature search was performed across four major bibliographic databases: Scopus, IEEE Xplore, ACM Digital

Library, and Web of Science. The search strategy used a combination of keywords and Boolean operators to capture relevant literature:

"Explainable AI" OR "XAI" OR "interpretable machine learning"
AND
"high-stakes" AND ("healthcare AI" OR "finance AI" OR "legal AI") AND
(evaluation OR regulation)

The search was restricted to peer-reviewed journal articles published between January 2015 and January 2025, capturing the most recent decade of XAI research across critical domains.

Inclusion criteria: Studies were eligible for inclusion if they met all the following criteria:

- Published in peer-reviewed journals
- Written in English
- Focused on healthcare, finance, or legal systems
- Provided a clear description of explainability methodology
- Included empirical validation or systematic framework analysis

Exclusion criteria: Studies were excluded if they:

- Were conference proceedings, editorials, or opinion articles
- Were non-peer-reviewed sources
- Lacked practical application or domain relevance

Study selection: The PRISMA-guided selection process ensured systematic filtering and transparency (Fig. 1):

- **Identification:** 1,284 records retrieved from databases
- **Duplicates removed:** 312
- **Screening:** 972 titles and abstracts screened; 726 excluded for irrelevance
- **Eligibility assessment:** 246 full-text articles assessed; 128 excluded due to:
 - Not belonging to a high-stakes domain (n = 49)
 - Absence of explainability method (n = 38)
 - Non-peer-reviewed (n = 21)
 - Insufficient methodological detail (n = 20)
- **Included:** 118 studies included in the qualitative synthesis

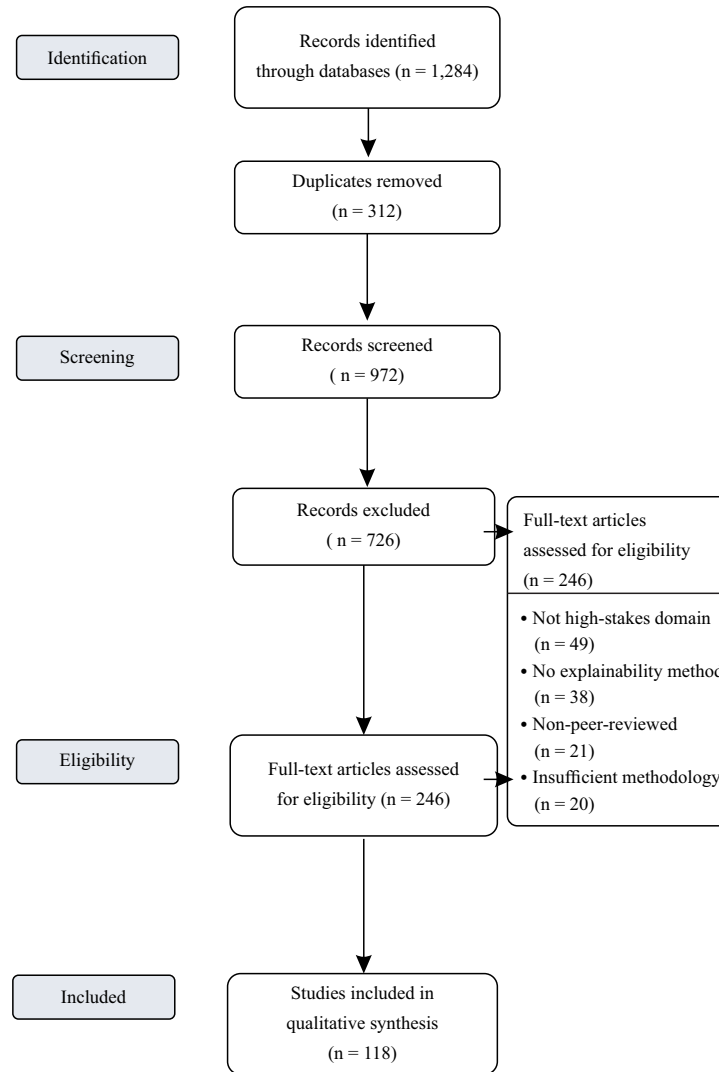


Fig. 1: PRISMA 2020 flow diagram depicting the systematic selection process of studies for inclusion in the review

Data extraction: From each study, the following variables were systematically extracted to enable structured analysis:

- **Domain:** Healthcare, finance, legal, or public sector
- **Model type:** Black-box or intrinsically interpretable
- **XAI method:** *post-hoc* or intrinsic interpretability approaches
- **Evaluation strategy:** Metrics such as fidelity, stability, and fairness
- **Human-centered validation:** Expert involvement, user studies, or cognitive assessment
- **Regulatory alignment:** Compliance with FDA, GDPR, EU AI Act, or other frameworks

This structured extraction facilitated cross-domain comparison of methodological trends, evaluation rigor, and

regulatory adherence. The approach ensures that the review not only summarizes the state-of-the-art but also identifies critical gaps in human-centered and regulatory-aligned XAI research.

The diagram illustrates the identification, screening, eligibility assessment, and final inclusion of 118 peer-reviewed studies on explainable AI in high-stakes domains.

RESULTS AND DISCUSSION

This systematic review synthesizes a decade of research (2015–2025) on explainable artificial intelligence (XAI) in high-stakes decision systems across healthcare, finance, legal, and public-sector domains. A total of 118 studies were analyzed, revealing clear domain patterns and methodological preferences.

Table 1: Evaluation metrics used across studies

Evaluation category	Frequency (n = 118)	Percentage
Fidelity metrics	87	74
Stability testing	52	44
Human-subject studies	32	27
Fairness metrics	41	35
Regulatory alignment discussion	40	34

Table 2: Distribution of studies by domain and XAI method

Domain	No. of studies	<i>post-hoc</i> (%)	Intrinsic (%)	Human-centered Eval (%)
Healthcare	57	63	37	31
Finance	38	58	42	24
Legal	18	67	33	22
Public sector	5	60	40	20

Domain distribution: Healthcare dominated the literature, representing 57 studies (48%), followed by finance with 38 studies (32%), legal systems with 18 studies (15%), and the public sector (non-clinical) with 5 studies (5%). The predominance of healthcare reflects the critical role of clinical decision support systems and the strong regulatory emphasis on patient safety. Finance-focused studies were primarily concerned with credit scoring, fraud detection, and risk management, whereas legal AI applications centered on algorithmic sentencing and risk assessment.

XAI Methodologies: A *post-hoc* interpretability methods were employed in 72 studies (61%), while 46 studies (39%) relied on intrinsically interpretable models.

***post-hoc* methods:** Common techniques included SHAP, LIME, counterfactual explanations, and saliency maps. Their strengths lie in being model-agnostic and compatible with deep learning architectures, allowing existing “black-box” models to be augmented with interpretability layers. However, limitations were evident: Explanations can be misleading or lack causal grounding, and their reliability for high-stakes decisions remains under scrutiny^{15,16}.

Intrinsic models: These included decision trees, rule-based systems, linear regression, and generalized additive models. Intrinsic models offer transparent reasoning and easier alignment with regulatory compliance. Nonetheless, performance trade-offs are often observed when applied to complex datasets, highlighting the tension between accuracy and interpretability^{17,18}.

Evaluation practices: Evaluation metrics were heterogeneous across studies. Fidelity was reported in 87 studies (74%), stability in 52 studies (44%), human-subject validation in 32 studies (27%), fairness metrics in 41 studies (35%), and

regulatory alignment in 40 studies (34%). Notably, only 32 studies directly involved domain experts in evaluating explanations, emphasizing a gap in human-centered assessment. Standardized benchmarks remain scarce, limiting cross-study comparability and complicating regulatory certification efforts. Human-centered evaluations, considering trust, cognitive load, and decision alignment, were underrepresented despite their centrality to high-risk applications as shown in Table 1.

Domain-specific insights¹⁹⁻²¹

Healthcare: Explainability is tightly linked to clinical trust, diagnostic validation, and patient safety. Saliency maps and attention mechanisms are frequently used to justify imaging-based predictions, yet concerns regarding their stability and fidelity persist. Intrinsic models such as decision trees and generalized additive models provide greater transparency but may reduce predictive performance, creating a trade-off that is particularly consequential in life-critical settings.

Finance: In financial applications, regulatory compliance drives XAI adoption. Credit scoring and fraud detection systems utilize feature attribution methods such as SHAP for transparency and accountability under GDPR and the EU AI Act. However, the interpretability of these explanations for non-technical stakeholders remains limited, and explanations can be unstable or manipulated.

Legal systems: Algorithmic sentencing and risk assessment tools face similar challenges. Opacity can undermine procedural fairness and due process, highlighting the importance of explainability in upholding justice and public trust shown in Table 2.

Regulatory integration: Forty studies (34%) explicitly discussed regulatory considerations. Healthcare studies referenced FDA guidance, while finance studies emphasized GDPR compliance. Legal AI studies are increasingly aligned with high-risk classifications under the EU AI Act. Despite growing regulatory pressures, “meaningful explanation” requirements remain technically and conceptually ambiguous, underscoring the need for interdisciplinary approaches to operationalize interpretability.

SYNTHESIS AND IMPLICATIONS

Across domains, *post-hoc* methods dominate due to reliance on complex models, yet intrinsic models remain essential where transparency is non-negotiable. Evaluation practices are fragmented, and human-centered assessment is insufficiently prioritized. The literature demonstrates a clear regulatory–technical gap: While laws increasingly mandate accountability, standardized, domain-sensitive frameworks for measuring explainability are lacking²²⁻²⁵.

Future research should develop hybrid approaches combining intrinsically interpretable architectures with selective *post-hoc* explanations, establish benchmark datasets for high-stakes evaluation, and incorporate robustness testing and longitudinal studies. Embedding XAI within socio-technical governance frameworks that align algorithmic design with ethical principles, regulatory mandates, and stakeholder expectations is critical. Such integration can ensure that XAI supports responsible, trustworthy AI deployment, particularly in domains where decisions have substantial human impacts^{26,27}.

LIMITATIONS

This review is subject to several limitations: First, publication bias toward English-language, peer-reviewed journals may exclude industrial or regulatory reports. Second, heterogeneity in study designs limited quantitative meta-analysis, necessitating narrative synthesis. Third, rapid advances in foundation models and generative AI emerging after 2023 may not be fully captured, presenting new explainability challenges related to multimodal reasoning and large-scale parameter opacity.

CONCLUSION

This systematic review evaluated Explainable Artificial Intelligence (XAI) approaches applied to high-stakes decision systems across healthcare, finance, legal, and public-sector

domains from 2015 to 2025. The findings indicate that *post-hoc* interpretability techniques such as SHAP, LIME, saliency maps, and counterfactual explanations remain dominant in practical applications, while inherently interpretable models are still underutilized despite their transparency advantages.

Explainability is largely driven by trust, accountability, and regulatory compliance rather than standardized technical benchmarks. Although some evaluation metrics are reported, the absence of unified standards and consistent human-centered validation limits comparability and weakens regulatory alignment.

A clear gap persists between increasing model complexity and the robustness of interpretability mechanisms. In high-risk contexts, explainability must be embedded within model design rather than applied as retrospective justification. Responsible AI deployment therefore depends on aligning robust interpretability methods, structured evaluation frameworks, and coherent regulatory guidance. Future work should focus on standardized benchmarks, domain-specific validation strategies, and hybrid models that balance performance with intrinsic transparency.

SIGNIFICANCE STATEMENT

This review highlights the critical need for structurally embedded explainability in high-stakes AI systems where decisions directly affect human lives and rights. By identifying gaps in evaluation standards and the overreliance on *post hoc* methods, it underscores the importance of harmonized benchmarks and human-centered validation. The findings provide a foundation for developing transparent, accountable, and regulation-ready AI frameworks that balance performance with meaningful interpretability across sensitive domains.

REFERENCES

1. Adadi, A. and M. Berrada, 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6: 52138-52160.
2. Lipton, Z.C., 2018. The mythos of model interpretability. *Commun. ACM*, 61: 36-43.
3. Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, Vol. 51. 10.1145/3236009.
4. Arrieta, A.B., N. Díaz-Rodríguez, J.D. Ser, A. Bennetot and S. Tabik *et al.*, 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58: 82-115.

5. Rizzo, M., A. Veneri, M. Marcuzzo, A. Zangari and A. Albarelli *et al.*, 2025. Machine learning models explanations as interpretations of evidence: A theoretical framework of explainability and its implications on high-stakes biomedical decision-making. *BMC Med. Res. Methodol.*, Vol. 25. 10.1186/s12874-025-02703-1.
6. Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267: 1-38.
7. Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1: 206-215.
8. Tjoa, E. and C. Guan, 2021. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Trans. Neural Networks Learn. Syst.*, 32: 4793-4813.
9. Mueller, S.T., E.S. Veinott, R.R. Hoffman, G. Klein, L. Alam, T. Mamun and W.J. Clancey, 2021. Principles of explanation in human-AI systems. *arXiv*, 10.48550/arXiv.2102.04972.
10. Mittelstadt, B.D., P. Allo, M. Taddeo, S. Wachter and L. Floridi, 2016. The ethics of algorithms: Mapping the debate. *Big Data Soc.*, Vol. 3. 10.1177/2053951716679679.
11. Wachter, S., B. Mittelstadt and L. Floridi, 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *Int. Data Privacy Law*, 7: 76-99.
12. Zou, X., W. Zhang and N. Zhao, 2025. From fragment to one piece: A review on AI-driven graphic design. *J. Imaging*, Vol. 11. 10.3390/jimaging11090289.
13. Mnih, V., K. Kavukcuoglu, D. Silver, A.A. Rusu and J. Veness *et al.*, 2015. Human-level control through deep reinforcement learning. *Nature*, 518: 529-533.
14. Samek, W., T. Wiegand and K.R. Müller, 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv*, 10.48550/arXiv.1708.08296.
15. Montavon, G., W. Samek and K.R. Müller, 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Process.*, 73: 1-15.
16. Mienye, I.D., G. Obaido, N. Jere, E. Mienye, K. Aruleba, I.D. Emmanuel and B. Ogbuokiri, 2024. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. *Inf. Med. Unlocked*, Vol. 51. 10.1016/j.imu.2024.101587.
17. Ghassemi, M., L. Oakden-Rayner and A.L. Beam, 2021. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Health*, 3: e745-e750.
18. Zhang, K., D. Wang, F. Lin, J. Xie and W. Zhou, 2026. A comprehensive review of explainable artificial intelligence in healthcare methods, evaluation, and clinical integration. *iScience*, Vol. 29. 10.1016/j.isci.2026.115026.
19. Rong, Y., T. Leemann, T.T. Nguyen, L. Fiedler and P. Qian *et al.*, 2024. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46: 2104-2122.
20. Lundberg, S.M., G. Erion, H. Chen, A. deGrave and J.M. Prutkin *et al.*, 2020. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2: 56-67.
21. Zhang, Q.S. and S.C. Zhu, 2018. Visual interpretability for deep learning: A survey. *Front. Inf. Technol. Electron. Eng.*, 19: 27-39.
22. Lundberg, S. and S.I. Lee, 2017. A unified approach to interpreting model predictions. *arXiv*, 10.48550/arXiv.1705.07874.
23. Biecek, P. and T. Burzykowski, 2021. *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. 1st Edn., Chapman and Hall/CRC, New York, USA, ISBN: 9780429027192, Pages: 324.
24. Jeon, E., W. Ko, J.S. Yoon and H.I. Suk, 2023. Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI. *IEEE Trans. Neural Networks Learn. Syst.*, 34: 739-749.
25. Leicht-Deobald, U., T. Busch, C. Schank, A. Weibel, S. Schafheitle, I. Wildhaber and G. Kasper, 2019. The challenges of algorithm-based HR decision-making for personal integrity. *J. Bus. Ethics*, 160: 377-392.
26. Kellogg, K.C., M.A. Valentine and A. Christin, 2020. Algorithms at work: The new contested terrain of control. *Acad. Manage. Ann.*, 14: 366-410.
27. Upadhyay, A.K. and K. Khandelwal, 2018. Applying artificial intelligence: Implications for recruitment. *Strategic HR Rev.*, 17: 255-258.