Journal of Artificial Intelligence and Computing Applications (2024) 2(1): 19–27

Critical perspective

The eXplainable Artificial Intelligence Paradox in Law: Technological Limits and Legal Transparency

Denise M. Trejo-Moncada
Universidad Nacional Autónoma de México, Facultad de Derecho

Keywords: explainability paradox, explainable artificial intelligence, artificial intelligence in law

PERSPECTIVE

The integration of Artificial Intelligence (AI) into legal systems offers transformative potential, promising enhanced efficiency and predictive accuracy. However, this progress also brings to the spotlight the explainability paradox: the unavoidable trade-off between the accuracy of complex Machine Learning (ML) and Deep Learning (DL) models and their lack of transparency. This paradox challenges foundational legal principles such as fairness, due process, and the right to explanation. While eXplainable AI (XAI) techniques have emerged to address this issue, their post-hoc nature, limited fidelity, and inaccessibility to non-expert stakeholders impede their practical utility in legal contexts. This paper critically reflects on the explainability paradox and its implications for AI-assisted legal decision-making, proposing a balanced framework to reconcile accuracy with transparency. By examining the limitations of current XAI methods and exploring the potential of inherently interpretable models, it highlights pathways to align AI systems with the procedural and ethical standards of the legal domain. These reflections not only address a gap in existing research but also challenge conventional reliance on opaque models, advocating for AI systems that prioritize trust, accountability, and legitimacy. This reflection invites interdisciplinary dialogue and encourages the development of AI tools that integrate technical performance with ethical and societal needs, ensuring the responsible adoption of AI in law.

Introduction

Artificial Intelligence (AI) has become an integral component of critical decision-making systems, transforming industries through its ability to analyze complex data and deliver predictive insights [1]. In fields such as healthcare, finance, and transportation, AI-driven systems have demonstrated exceptional performance, enabling faster and more accurate decisions [2, 3]. The legal sector, traditionally reliant on human expertise and interpretability, is increasingly exploring the adoption of AI technologies, particularly Machine Learning (ML) and Deep Learning (DL), to enhance processes such as case management, risk assessment, and evidence analysis [4, 5]. These models, known for their ability to uncover patterns and predict outcomes with unprecedented precision, promise significant advancements in efficiency and accuracy. However, this progress comes at a cost: the inherent opacity of complex ML and DL models raises critical concerns about explainability, trust, and accountability, particularly in contexts where transparency is a foundational requirement [6, 4].

This rapid adoption of AI in legal contexts introduces a fundamental paradox: while the legal system is built on principles of transparency, accountability, and fairness, modern ML and DL models often operate as black boxes [7, 8]. These models generate predictions and decisions with exceptional accuracy, yet their inner workings remain largely opaque, even to their developers [9, 10]. Legal decisions—whether related to sentencing, risk assessment, or evidence evaluation—demand reasoning that is not only accurate but also explainable and justifiable to all stakeholders [11, 12]. In this setting, the lack of interpretability in AI systems conflicts directly with the legal requirement for clear, understandable decision-making. This tension raises critical questions: Can AI systems truly achieve the level of transparency necessary for legal legitimacy, or does their reliance on complexity undermine trust and fairness? [10, 7]

In response to the opacity of modern AI systems, the emerging field of eXplainable AI (XAI) seeks to address the interpretability challenge by providing insights into how complex ML and DL models generate their outputs. Techniques such as SHAP (Shapley Additive Explanations), LIME (Local

E-mail address: denise.trejo@aaaimx.org
https://doi.org/10.5281/zenodo.14692066

Interpretable Model-agnostic Explanations), and visualization methods like Grad-CAM have been developed to offer post-hoc explanations, enabling a degree of transparency in otherwise black-box systems [13, 14]. These tools attempt to highlight the relationships between inputs and predictions, offering approximations of the model's reasoning process [15, 16]. However, despite these advancements, XAI techniques remain limited: they often provide partial or oversimplified explanations that fail to capture the true internal logic of complex models [17, 18]. Furthermore, these explanations are frequently difficult for non-experts to interpret, raising concerns about their adequacy for legal systems where transparency, clarity, and trust are paramount [13, 18]. This gap highlights the ongoing struggle to align the technical capabilities of XAI with the rigorous standards of transparency required in legal decision-making.

This paper critically reflects on the limitations of XAI within the context of legal applications, where transparency and accountability are non-negotiable principles. While ML and DL models offer unprecedented predictive capabilities, their lack of interpretability creates a significant barrier to their adoption in legal systems. This work argues that the paradox between explainability and performance must be addressed to ensure that AI tools align with legal requirements for transparency, fairness, and trust. Without overcoming this challenge, the deployment of AI in sensitive legal contexts risks undermining the very foundations of legitimacy and justice that the legal system upholds. Moving forward, achieving a balance between model performance and explainability must become a central focus for researchers and practitioners seeking to integrate AI into legal decision-making processes responsibly.

Position

The increasing adoption of AI in legal systems brings to the forefront a critical paradox: the tension between accuracy and explainability in ML and DL models [19, 20]. On one hand, the unparalleled predictive power of complex AI systems enables them to identify patterns and deliver insights that surpass human capabilities [21, 22]. On the other hand, the very complexity that drives their performance renders these models opaque and difficult to interpret, even for their creators [23, 24]. In legal contexts, where decisions must be transparent, justifiable, and open to scrutiny, this lack of interpretability presents a significant barrier [25]. However, it is worth noting that human decision-making in legal processes is not without flaws, as biases, inconsistencies, or even corruption can occasionally compromise fairness and accountability. AI systems, despite their opacity, offer the potential for greater objectivity and the ability to inspect and analyze decision-making processes retrospectively. The challenge lies in reconciling the need for high-performing AI systems with the equally important requirement for clear and understandable reasoning—a fundamental pillar of fairness and trust within legal processes [26]. Addressing this paradox is not merely a technical necessity but a critical step toward ensuring the ethical and legitimate integration of AI into the legal domain.

At the core of the explainability paradox is a well-recognized trade-off in AI: simpler, interpretable models often sacrifice accuracy, while complex, high-performing models, such as deep neural networks, sacrifice transparency [27, 28]. Interpretable models—such as decision trees, linear regression, and rule-based systems—are inherently easier to understand and explain [29, 30]. These models allow stakeholders to trace decisions back to specific inputs, offering a level of clarity that aligns with the legal system's need for justification and accountability [31, 32]. However, their simplicity limits their ability to capture intricate patterns within large, complex datasets, often resulting in lower predictive accuracy [33]. Case-Based Reasoning (CBR) offers an alternative approach to bridging this gap by retrieving and reasoning through similar past cases [34]. Its interpretable framework ensures that decisions are supported by precedent [35], which aligns naturally with the legal system's emphasis on contextual and historical consistency.

In contrast, DL models excel at delivering superior performance by leveraging massive datasets and complex architectures to uncover subtle relationships that are imperceptible to traditional approaches [33, 36]. Yet, this complexity comes at the cost of interpretability: the reasoning behind a model's prediction remains opaque, hidden within layers of parameters and computations that defy human understanding [37, 38]. While CBR systems may not achieve the predictive power of DL models, they can enhance interpretability by explicitly tying new decisions to past examples, providing a transparent foundation for explanations. However, it is important to recognize that even opaque DL models may offer opportunities for retrospective inspection and analysis, capabilities that human decision-making processes often lack. In legal applications—where decisions affect lives, freedoms, and rights—accuracy alone is insufficient. The law demands not only correct outcomes but also explanations that can be understood, challenged, and justified [27, 39]. While the opacity of AI systems raises concerns, they may still serve as valuable tools to complement human decision-making, particularly in contexts where human biases or a lack of accountability have historically undermined trust. This trade-off highlights a fundamental challenge in aligning AI's technical capabilities with the ethical and procedural standards required in legal systems [40].

In response to the opacity of complex ML and DL models, significant efforts have been made in developing XAI techniques to make their decision-making processes more interpretable. Methods such as SHAP, LIME, Feature Importance (FI), and visualization tools like Grad-CAM have emerged as widely adopted solutions [13, 18, 16]. These techniques attempt to provide insights into how input features influence model outputs, offering a degree of transparency that was previously unattainable [41, 42]. However, despite these advancements, current XAI methods remain limited in key ways that significantly impact their suitability for legal applications [43, 44].

First, most XAI techniques are post-hoc in nature, meaning they generate explanations after the model has produced its output, rather than ensuring inherent transparency within the model itself [45, 46]. This creates a reliance on approximations, which may not accurately reflect the true reasoning process of the underlying model [47, 48]. Second, the explanations provided by XAI methods are often oversimplified, reducing complex interactions into digestible summaries that can be incomplete or even misleading [49, 8]. For instance, FI scores or Grad-CAM's heatmaps may highlight correlations but fail to convey the relationships or causal factors driving predictions [50]. Lastly, the accessibility of these explanations remains a significant barrier. While XAI outputs may be interpretable for AI experts, they are often too technical or abstract for non-expert stakeholders, such as judges, lawyers, or defendants [51, 52]. In a legal setting—where transparency must be both accurate and comprehensible—these limitations hinder the practical usability of XAI, undermining its ability to meet the rigorous standards of justification and accountability required by the law.

The use of black-box AI systems in legal contexts raises profound implications for fundamental legal principles such as due process, fairness, and the right to explanation [11, 53]. At the heart of the legal system lies the requirement for decisions to be transparent, justifiable, and open to scrutiny [54, 55]. When decisions are influenced or determined by opaque AI models—whose reasoning cannot be fully understood or explained—it becomes challenging for stakeholders to evaluate whether those decisions are fair, unbiased, or free from error [10]. However, it is also worth noting that human decision-makers, despite their interpretability, can at times be equally opaque—whether due to cognitive limitations, unconscious biases, or intentional withholding of reasoning. This parallel suggests that while black-box AI systems introduce challenges, they also provide an opportunity for systematic inspection and reproducibility that human decisions often lack. By documenting decision-making processes through algorithms, AI can offer a framework for identifying errors or biases retrospectively, fostering a level of accountability that is not always achievable in human-driven systems. Addressing the transparency issues inherent in AI is crucial, but leveraging these tools to complement human decision-making could mitigate long-standing concerns about fairness and consistency in the legal domain [56].

Furthermore, the deployment of black-box systems threatens to erode trust—not only in the AI tools themselves but also in the legal institutions adopting them. Trust in the justice system is built on its ability to deliver outcomes that are not only accurate but also explainable and consistent. Opaque AI models, by their nature, introduce uncertainty and raise doubts about accountability, especially when errors occur or biases emerge. However, trust in human decision-making is not always guaranteed either, as it can be compromised by biases, inconsistencies, or even intentional misconduct. AI systems, despite their opacity, offer unique opportunities for post-hoc analysis and continuous improvement, allowing stakeholders to retrospectively evaluate and refine decision-making processes in ways that are often impossible with human decisions. While such systems risk creating a perception that justice is being "outsourced" to inscrutable algorithms, their ability to document and analyze decisions systematically presents a pathway to enhance transparency and accountability if implemented responsibly. Balancing these perspectives is essential to maintaining the legitimacy of legal processes.

Without addressing the explainability problem, the widespread deployment of black-box AI systems could lead to significant ethical and legal challenges, including violations of fairness, potential misuse, and unintended harm. These challenges expose the urgent need for a careful, measured approach to integrating AI in legal decision-making—one that prioritizes transparency and accountability over blind reliance on performance. Failure to resolve these issues may not only weaken confidence in AI but also jeopardize the integrity of the legal system itself.

To bridge the gap between AI's predictive capabilities and the legal system's demand for transparency, it is evident that current solutions remain insufficient. While hybrid approaches and inherently interpretable AI models hold promise, significant technical and ethical challenges persist. Achieving a balance between accuracy and transparency will require ongoing research, responsible design, and a commitment to maintaining human oversight in AI-assisted legal processes. However, it is equally important to recognize AI's potential to complement the legal system by offering consistency and reducing biases that can affect human decision-making. By leveraging AI's ability to document and standardize decision-making processes, legal institutions have an opportunity to evolve toward more equitable and objective outcomes. Until such advancements are realized, the widespread deployment of opaque systems in legal contexts risks undermining trust, accountability, and the very principles the legal system seeks to uphold.

Discussion

The explainability paradox—where increased accuracy in AI systems comes at the cost of transparency—has far-reaching implications for legal systems attempting to integrate ML and DL models. At its core, this paradox challenges the foundational principles of the law, which rely on decisions that are transparent, justifiable, and contestable [57, 58]. Legal systems are not merely tasked with reaching accurate outcomes but must also ensure that decisions can be understood and trusted by all stakeholders, including judges, lawyers, defendants, and the public [59, 60]. While the opacity of AI models can undermine these key pillars of justice, it is essential to recognize that human decision-making is not inherently more transparent. Cognitive biases, inconsistencies, and even deliberate misconduct can obscure the reasoning behind human judgments. In contrast, AI systems provide systematic documentation and opportunities for retrospective analysis, offering a mechanism for identifying errors or biases that might otherwise remain hidden. By leveraging these strengths, AI has the potential to address some of the long-standing challenges associated with human decision-making while complementing traditional legal frameworks.

Moreover, the paradox exposes a deeper ethical tension: efficiency versus legitimacy. While AI systems promise increased efficiency by automating processes and enhancing predictive accuracy, their opacity risks eroding public trust in legal institutions [61, 62]. For example, AI-based risk assessment tools used in sentencing or parole decisions may produce accurate predictions, but without clear explanations, stakeholders cannot assess whether these decisions are fair or free from bias [63]. However, human decision-making is not immune to similar challenges; biases, inconsistencies, and even corruption can sometimes compromise fairness and accountability in legal judgments. In this context, AI systems offer a unique opportunity to mitigate such issues by providing systematic, data-driven insights that are less prone to individual bias and more amenable to post-hoc inspection. Nevertheless, the adoption of opaque AI systems—without addressing their explainability limitations—risks creating a perception that justice is being outsourced to inscrutable algorithms, thereby undermining the legitimacy and credibility of legal outcomes [64]. This tension highlights the urgent need for a more responsible approach to integrating AI into the legal system, one that not only prioritizes technical performance but also leverages AI's potential to enhance fairness while upholding the ethical standards upon which the rule of law is built.

Current efforts in XAI, such as SHAP, LIME, and Grad-CAM, have made progress in shedding light on black-box predictions by offering insights into feature importance and decision boundaries. While these techniques improve transparency, they remain insufficient for legal applications, where clarity, reliability, and accessibility are non-negotiable.

A major limitation of XAI lies in its post-hoc nature—generating explanations after predictions are made. These approximations may fail to reflect the model's true reasoning, raising concerns about their fidelity in sensitive legal decisions. Additionally, XAI outputs are often oversimplified, reducing complex relationships to summaries that can obscure critical connections and mislead stakeholders.

Finally, XAI explanations, such as feature scores or heatmaps, are often too technical for non-expert stakeholders like judges or lawyers. Legal applications require explanations that are clear, actionable, and comprehensible to ensure decisions can be scrutinized and justified [65, 52]. These challenges highlight the need for further advancements in inherently interpretable models that align AI systems with the rigorous standards of fairness and accountability required by the legal system [66, 67].

Hybrid models, which combine interpretable algorithms with the predictive power of black-box systems [68], offer a promising middle ground to address the explainability paradox. For example, interpretable models like decision trees or linear regression could handle critical decisions requiring justification, while complex DL models can assist in auxiliary tasks or pre-processing stages. This layered approach maintains transparency where it matters most while leveraging the accuracy of advanced AI systems. However, hybrid solutions are not without limitations. Seamlessly integrating interpretable and black-box components remains a technical challenge, as discrepancies between models could lead to inconsistencies. Additionally, determining which stages require human oversight versus AI-driven decisions introduces added complexity. Despite these challenges, hybrid models represent a feasible pathway for balancing accuracy and transparency, particularly in legal contexts where trust and accountability are paramount.

A more sustainable solution to the explainability paradox lies in developing inherently interpretable AI models—systems designed for transparency from the outset. Unlike post-hoc explanations, these models integrate interpretability into their structure, ensuring that decision-making processes are both clear and traceable [36]. Recent progress in techniques such as Generalized Additive Models (GAMs), explainable neural networks, and sparse models demonstrates that transparency and performance can coexist, particularly in tasks with structured data [69, 70].

Inherently interpretable models—designed for transparency from the outset—offer a promising solution to the explainability paradox. These models integrate interpretability into their structure, ensuring that decision-making processes are clear and traceable [36]. While such models currently lag behind DL systems in handling unstructured, high-dimensional data, they can mitigate many risks associated with black-box AI in legal contexts. By prioritizing models that are understandable by design, stakeholders gain greater confidence in AI outputs, ensuring alignment with legal standards of fairness and accountability. However, even black-box systems have advantages: their systematic documentation of decision-making pathways enables retrospective analysis and error detection, which are often lacking in human-driven processes. Advancing both inherently interpretable models and strategies to enhance the transparency of complex systems is essential for achieving a balance between accuracy and explainability while addressing the limitations of both human and AI decision-making.

While AI systems offer valuable support in legal decision-making, human oversight remains indispensable. Legal processes require contextual understanding, ethical judgment, and adherence to evolving societal standards—capabilities that AI, regardless of its sophistication, cannot fully replicate [71, 72]. However, AI systems, even when opaque, provide a systematic approach to decision-making that can mitigate human biases and inconsistencies. By combining AI's ability to process large datasets and document decision-making pathways with human expertise in applying broader reasoning, stakeholders can ensure that final decisions remain accountable, fair, and aligned with legal principles. This synergy reduces the risks of relying solely on either human or AI decision-making, creating a collaborative framework where AI enhances consistency and trust while humans provide essential oversight.

The collaborative use of AI and human oversight in legal systems has implications beyond the legal domain, offering a roadmap for ethical AI adoption in other high-stakes areas such as healthcare, finance, and governance. Lessons learned from aligning AI with legal standards—such as prioritizing transparency, fairness, and accountability—can inform the responsible deployment of AI across industries. Future research must focus on developing inherently interpretable models that handle unstructured, high-dimensional data without sacrificing performance, as well as refining hybrid approaches to strike a balance between accuracy and transparency. Additionally, interdisciplinary collaboration between AI researchers, legal scholars, and ethicists will be critical to ensuring that AI systems are not only technically robust but also ethically aligned with societal values. By addressing these challenges, AI can evolve into a tool that enhances decision-making processes while maintaining trust and accountability in sensitive applications.

Resolving the explainability paradox has implications beyond the legal domain, offering a roadmap for AI adoption in other high-stakes areas like healthcare, finance, and governance. Lessons learned from aligning AI with legal standards—such as prioritizing transparency, fairness, and accountability—can inform ethical AI practices across industries. Future research must focus on developing inherently interpretable models capable of handling complex, unstructured data without sacrificing performance, as well as refining hybrid approaches to strike a balance between accuracy and explainability.

Additionally, interdisciplinary collaboration between AI researchers, legal scholars, and ethicists will be critical to ensuring AI systems are not only technically robust but also ethically aligned with societal values. By addressing these challenges, AI can evolve into a tool that enhances decision-making processes while maintaining trust and accountability in sensitive applications.

Conclusions

This paper has examined the explainability paradox in AI: the tension between the exceptional predictive performance of ML and DL models and their lack of transparency. While these models hold the potential to revolutionize legal systems, their opacity raises significant challenges for ensuring that decisions remain transparent, justifiable, and accountable. This paradox underscores the critical need to align AI tools with legal principles, ensuring their integration enhances the integrity and fairness of legal processes rather than undermining them.

Current XAI techniques, though instrumental in improving transparency, remain limited by their post-hoc nature, oversimplification of complex model behavior, and inaccessibility to non-experts. These limitations are particularly pronounced in legal contexts, where clarity and justification are paramount. To address this, hybrid models that combine interpretable algorithms with the predictive power of black-box systems offer a practical interim solution. In the longer term, the development of inherently interpretable AI systems, designed for transparency without sacrificing performance, represents a more sustainable path forward. Crucially, maintaining human oversight remains indispensable to uphold fairness, accountability, and public trust.

Moving forward, addressing the explainability paradox requires a collaborative effort between researchers, developers, and legal practitioners. Future work should prioritize advancing inherently interpretable models, refining XAI techniques to enhance fidelity and accessibility, and fostering interdisciplinary collaboration between AI experts, legal scholars, and ethicists. By aligning technological advancements with ethical and legal standards, we can harness AI's potential to complement human judgment, mitigate biases, and strengthen the principles of fairness and trust that underpin the legal system.

CRediT authorship contribution statement

Denise Trejo-Moncada: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used ChatGPT in order to improve readability. After using this tool, the author reviewed and edited the content as needed and took full responsibility for the content of the publication.

Declaration of competing interest

The author declares that she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author would like to express her deepest gratitude to F.E. Heart for the invaluable discussions that greatly enhanced the depth of this work. Special thanks also go to D.P. Assenger, whose insightful feedback during the revision process was instrumental in refining the manuscript. Their contributions have been truly appreciated.

References

[1] M. Elhaddad and S. Hamam, “Ai-driven clinical decision support systems: An ongoing pursuit of potential,” *Cureus*, vol. 16, 2024.

[2] M. Jeyaraman, S. Balaji, N. Jeyaraman, and S. Yadav, “Unraveling the ethical enigma: Artificial intelligence in healthcare,” *Cureus*, vol. 15, 2023.

[3] K. Yekaterina, “Challenges and opportunities for ai in healthcare,” *International Journal of Law and Policy*, 2024.

[4] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, “Trustworthy artificial intelligence: A review,” *ACM Computing Surveys (CSUR)*, vol. 55, pp. 1–38, 2022.

[5] V. Lai, C. Chen, A. Smith-Renner, Q. Liao, and C. Tan, “Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[6] N. Biller-Andorno, A. Ferrario, S. Joebges *et al.*, “Ai support for ethical decision-making around resuscitation: proceed with care,” *Journal of Medical Ethics*, vol. 48, pp. 175–183, 2020.

[7] N. Thalpage, “Unlocking the black box: Explainable artificial intelligence (xai) for trust and transparency in ai systems,” *Journal of Digital Art & Humanities*, 2023.

[8] G. Chaudhary, “Explainable artificial intelligence (xai): Reflections on judicial system,” *Kutafin Law Review*, 2024.

[9] M. S. M. d. Encarnacao, M. Anastasiadou, and V. Santos, “Framework for the application of explainable artificial intelligence techniques in the service of democracy,” *Transforming Government: People, Process and Policy*, 2024.

[10] V. Gupta, S. Shukla, and K. Nikita, “Cracking the code: Enhancing trust in ai through explainable models,” *resmilitaris*, 2024.

[11] M. T. Sacramed, “Reviewing the philippines legal landscape of artificial intelligence (ai) in business: Addressing bias, explainability, and algorithmic accountability,” *International Journal of Research and Innovation in Social Science*, 2024.

[12] M. S. Marques da Encarnacao, M. Anastasiadou, and V. Santos, “Framework for the application of explainable artificial intelligence techniques in the service of democracy,” *Transforming Government: People, Process and Policy*, 2024.

[13] L. Zou, H. Goh, C. Liew, J. Quah, G. T. Gu, J. J. Chew, M. P. Kumar, C. Ang, and A. W. A. Ta, “Ensemble image explainable ai (xai) algorithm for severe community-acquired pneumonia and covid-19 respiratory infections,” *IEEE Transactions on Artificial Intelligence*, vol. 4, pp. 242–254, 2023.

[14] H. Byeon, “Advances in machine learning and explainable artificial intelligence for depression prediction,” *International Journal of Advanced Computer Science and Applications*, 2023.

[15] A. G, S. B. Madagaonkar, and R. C. H, “Unveiling the black box: A comprehensive review of explainable ai techniques,” *International Journal of Scientific Research in Engineering and Management*, 2024.

[16] F. A. Undie, L. V. Kruglova, M. O. Okache, V. A. Undie, and R. A. Aloye, “Exploring explainable artificial intelligence (xai) to enhance healthcare decision support systems in nigeria,” *Journal of Innovative Research*, 2024.

[17] M. Saarela and V. Podgorelec, “Recent applications of explainable ai (xai): A systematic literature review,” *Applied Sciences*, 2024.

[18] F. Abdullakutty, Y. Akbari, S. Al-Maadeed, A. Bouridane, I. M. Talaat, and R. Hamoudi, “Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis,” *Frontiers in Medicine*, 2024.

[19] S. Veer, L. Riste, S. Cheraghi-Sohi *et al.*, “Trading off accuracy and explainability in ai decision-making: findings from 2 citizens’ juries,” *Journal of the American Medical Informatics Association*, vol. 28, pp. 2128–2138, 2021.

[20] K. Atkinson, T. J. M. Bench-Capon, and D. Bollegala, “Explanation in ai and law: Past, present and future,” *Artificial Intelligence*, vol. 289, p. 103387, 2020.

[21] G. Chaudhary, “Explainable artificial intelligence (xai): Reflections on judicial system,” *Kutafin Law Review*, 2024.

[22] R. Ejjami, “Ai-driven justice: Evaluating the impact of artificial intelligence on legal systems,” *International Journal For Multidisciplinary Research*, 2024.

[23] G. Vilone and L. Longo, “Notions of explainability and evaluation approaches for explainable artificial intelligence,” *Information Fusion*, vol. 76, pp. 89–106, 2021.

[24] W. Guo, "Explainable artificial intelligence for 6g: Improving trust between human and machine," *IEEE Communications Magazine*, vol. 58, pp. 39–45, 2020.

[25] G. Joshi, "A systematic review on explainable ai in legal domain," *International Journal for Research in Applied Science and Engineering Technology*, 2024.

[26] T. Ha, S. Lee, and S. Kim, "Designing explainability of an artificial intelligence system," in *Proceedings of the Technology, Mind, and Society*, 2018.

[27] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 22 071–22 080, 2019.

[28] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, 2020.

[29] M. Nauta, J. Trienes, S. Pathak *et al.*, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, 2022.

[30] C. P. R. Vieira and L. A. Digiampietri, "Machine learning post-hoc interpretability: A systematic mapping study," *XVIII Brazilian Symposium on Information Systems*, 2022.

[31] M. M. Xu, P. Watkinson, and T. Zhu, "Explainable ai for clinical risk prediction: A survey of concepts, methods, and modalities," *ArXiv*, vol. abs/2308.08407, 2023.

[32] K. Sankaran, "Data science principles for interpretable and explainable ai," *ArXiv*, vol. abs/2405.10552, 2024.

[33] Q. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 27–39, 2018.

[34] J. A. Recio-García, H. Parejas-Llanovarced, M. G. Orozco-del Castillo, and E. E. Brito-Borges, "A case-based approach for the selection of explanation algorithms in image classification," in *Case-Based Reasoning Research and Development: 29th International Conference, ICCBR 2021, Salamanca, Spain, September 13–16, 2021, Proceedings 29*. Springer, 2021, pp. 186–200.

[35] M. G. Orozco-del Castillo, J. A. Recio-Garcia, and E. C. Orozco-del Castillo, "Item-specific similarity assessments for explainable depression screening," in *International Conference on Case-Based Reasoning*. Springer, 2024, pp. 430–444.

[36] A. Somani, A. Horsch, A. Bopardikar, and D. K. Prasad, "Propagating transparency: A deep dive into the interpretability of neural networks," *Nordic Machine Intelligence*, 2024.

[37] X. Li, H. Xiong *et al.*, "Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond," *Knowledge and Information Systems*, vol. 64, pp. 3197–3234, 2021.

[38] L. Munroe, M. d. Silva *et al.*, "Applications of interpretable deep learning in neuroimaging: A comprehensive review," *Imaging Neuroscience*, vol. 2, pp. 1–37, 2024.

[39] A. M. Hanif, S. Beqiri, P. Keane, and J. Campbell, "Applications of interpretability in deep learning models for ophthalmology," *Current Opinion in Ophthalmology*, vol. 32, pp. 452–458, 2021.

[40] F. Fan, J. Xiong, M. Li, and G. Wang, "On interpretability of artificial neural networks: A survey," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, pp. 741–760, 2020.

[41] A. G, S. B. Madagaonkar, and R. C. H, "Unveiling the black box: A comprehensive review of explainable ai techniques," *International Journal of Scientific Research in Engineering and Management*, 2024.

[42] D. Tang, J. Chen, L. Ren, X. Wang, D. Li, and H. Zhang, "Reviewing cam-based deep explainable methods in healthcare," *Applied Sciences*, 2024.

[43] N. Y. Fares, D. Nedeljkovic, and M. Jammal, "Ai-enabled iot applications: Towards a transparent governance framework," *2023 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, pp. 109–114, 2023.

[44] F. Xu, L. Jiang, W. He, G. Huang, Y. Hong, F. Tang, J. Lv, Y. Lin, Y. Qin, R. Lan, X. Pan, S. Zeng, M. Li, Q. Chen, and N. Tang, "The clinical value of explainable deep learning for diagnosing fungal keratitis using in vivo confocal microscopy images," *Frontiers in Medicine*, vol. 8, 2021.

[45] A. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review," *Applied Sciences*, vol. 11, p. 5088, 2021.

[46] B. M. de Vries, G. Zwezerijnen, G. Burchell, F. V. van Velden, C. W. M. van der Houven van Oordt, and R. Boellaard, "Explainable artificial intelligence (xai) in radiology and nuclear medicine: a literature review," *Frontiers in Medicine*, vol. 10, 2023.

[47] M. Mainali and R. O. Weber, "What's meant by explainable model: A scoping review," *ArXiv*, vol. abs/2307.09673, 2023.

[48] C. P. R. Vieira and L. A. Digiampietri, “Machine learning post-hoc interpretability: a systematic mapping study,” in *XVIII Brazilian Symposium on Information Systems*, 2022.

[49] M. Fontes, J. D. S. D. Almeida, and A. Cunha, “Application of example-based explainable artificial intelligence (xai) for analysis and interpretation of medical imaging: A systematic review,” *IEEE Access*, vol. 12, pp. 26 419–26 427, 2024.

[50] M. T. Keane and E. M. Kenny, “How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems,” in *Lecture Notes in Computer Science*, 2019, pp. 155–171.

[51] M. Velmurugan, C. Ouyang, Y. Xu, R. Sindhgatta, B. Wickramanayake, and C. Moreira, “Developing guidelines for functionally-grounded evaluation of explainable artificial intelligence using tabular data,” in *2024 International Conference on Explainable AI*, 2024.

[52] Z. Zhou, M. Hu, M. Salcedo, N. Gravel, W. Yeung, A. Venkat, D. Guo, J. Zhang, N. Kannan, and S. Li, “Xai meets biology: A comprehensive review of explainable ai in bioinformatics applications,” *ArXiv*, vol. abs/2312.06082, 2023.

[53] P. N. Thakre, P. R. Sahu, P. K. Soni, D. Bisen, and A. Sahu, “Evaluating transparency and explainability in ai-driven planning and scheduling: A comprehensive literature review,” *International Journal of Innovative Research in Science, Engineering and Technology*, 2023.

[54] M. I. Konkov, “Ethical issues of implementing artificial intelligence in medicine,” *Digital Diagnostics*, 2023.

[55] G. Chaudhary, “Explainable artificial intelligence (xai): Reflections on judicial system,” *Kutafin Law Review*, 2024.

[56] K. Ingram, “Ai and ethics: Shedding light on the black box,” *International Review of Information Ethics*, 2020.

[57] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, “Explainable, trustworthy, and ethical machine learning for healthcare: A survey,” *Computers in biology and medicine*, vol. 149, p. 106043, 2021.

[58] A. Kale, T. C. Nguyen, F. Harris, C. Li, J. Zhang, and X. Ma, “Provenance documentation to enable explainable and trustworthy ai: A literature review,” *Data Intelligence*, vol. 5, pp. 139–162, 2022.

[59] U. Bhatt, Y. Zhang, J. Antorán, Q. Liao, P. Sattigeri, R. Fogliato, G. G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, A. Weller, and A. Xiang, “Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty,” *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2020.

[60] M. Hossain, S. Das, B. Krishnamurthy, and S. G. Shiva, “Explainability of artificial intelligence systems: A survey,” *2023 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6, 2023.

[61] B. C. Cheong, “Transparency and accountability in ai systems: safeguarding wellbeing in the age of algorithmic decision-making,” *Frontiers in Human Dynamics*, 2024.

[62] G. B. Mensah, “Artificial intelligence and ethics: A comprehensive reviews of bias mitigation,transparency, and accountability in ai systems,” *Africa Journal For Regulatory Affairs*, 2024.

[63] E. E. Agu, A. O. Abhulimen, A. N. Obiki-Osafiele, O. S. Osundare, I. A. Adeniran, and C. P. Efunniyi, “Discussing ethical considerations and solutions for ensuring fairness in ai-driven financial services,” *International Journal of Frontline Research in Multidisciplinary Studies*, 2024.

[64] S. Akter, “Ethical ai development for sustainable enterprises: A review of integrating responsible ai with iot and enterprise systems,” *Journal of Artificial Intelligence General Science*, 2024.

[65] E. Owens, B. Sheehan, M. Mullins, M. Cunneen, J. Ressel, and G. Castignani, “Explainable artificial intelligence (xai) in insurance,” *Risks*, 2022.

[66] S. Alam and Z. Altıparmak, “Xai-cf - examining the role of explainable artificial intelligence in cyber forensics,” *ArXiv*, vol. abs/2402.02452, 2024.

[67] A. Kalyakulina and I. Yusipov, “explainable artificial intelligence (xai) in age prediction: A systematic review,” *ArXiv*, vol. abs/2307.13704, 2023.

[68] G. Q. Álvarez, M. J. del Jesús Díaz, and P. G. García, “Explainable artificial intelligence: An overview on hybrid models,” in *Proceedings of the First Multimodal, Affective and Interactive eXplainable AI Workshop (MAI-XAI24 2024), co-located with 27th European Conference on Artificial Intelligence (ECAI 2024)*, J. M. Alonso-Moral, Z. Anthis, R. Berlanga, A. Catalá, P. Cimiano, P. Flach, E. Hüllermeier, T. Miller, O. Mitruţ, D. Mindlin, G. Moise, A. Moldoveanu, F. Moldoveanu, K. Sokol, and A. Soroa, Eds. Santiago de Compostela, Spain: CEUR Workshop Proceedings, 2024, pp. 49–60. [Online]. Available: http://ceur-ws.org/Vol-3803/

[69] C. Lee, M. Samad, I. Hofer, M. Cannesson, and P. Baldi, “Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality,” *NPJ Digital Medicine*, vol. 4, 2021.

[70] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” *Entropy*, vol. 23, 2020.

[71] M. Yurrita, T. Draws, A. Balayn, D. Murray-Rust, N. Tintarev, and A. Bozzon, “Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[72] K. S. Reddy, M. Kethan, S. M. Basha, A. Singh, P. Kumar, and D. Ashalatha, “Ethical and legal implications of ai on business and employment: Privacy, bias, and accountability,” in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, vol. 1, 2024, pp. 1–6.