6-2024

Navigating the Complexities of AI: The Critical Role of Interpretability and Explainability in Ensuring Transparency and Trust

Emily Barnes

James Hutson

International Journal of Multidisciplinary and Current Educational Research (IJMCER)
ISSN: 2581-7027 ||Volume|| 6 ||Issue|| 3 ||Pages 248-256 ||2024||

Navigating the Complexities of AI: The Critical Role of Interpretability and Explainability in Ensuring Transparency and Trust

[1,]Emily Barnes, EdD, PhD, [2,]James Hutson, PhD
[1,]*Capitol Technology University*
https://orcid.org/0000-0001-9401-0186
[2,]*Lindenwood University*
https://orcid.org/0000-0002-0578-6052

ABSTRACT: The interpretability and explainability of deep neural networks (DNNs) are paramount in artificial intelligence (AI), especially when applied to high-stakes fields such as healthcare, finance, and autonomous driving. The need for this study arises from the growing integration of AI into critical areas where transparency, trust, and ethical decision-making are essential. This paper explores the impact of architectural design choices on DNN interpretability, focusing on how different architectural elements like layer types, network depth, connectivity patterns, and attention mechanisms affect model transparency. Methodologically, the study employs a comprehensive review of case studies and experimental results to analyze the balance between performance and interpretability in DNNs. It examines real-world applications to demonstrate the importance of interpretability in sectors like healthcare, finance, and autonomous driving. The study also reviews practical tools such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to assess their effectiveness in enhancing model transparency. The results underscore that interpretability facilitates better decision-making, accountability, and compliance with regulatory standards. For instance, using SHAP in environmental monitoring helps policymakers understand the key drivers of air quality, leading to informed interventions. In education, LIME aids educators in personalizing learning by highlighting factors influencing student performance. The findings also reveal that incorporating attention mechanisms and hybrid model architectures can significantly improve interpretability without compromising performance.

KEYWORDS: Interpretability, Explainability, Deep neural networks, AI transparency

I. INTRODUCTION

As artificial intelligence (AI) systems become increasingly integrated into various aspects of society, the demand for transparency in their operations grows correspondingly. Interpretability and explainability are two crucial facets of AI that address this need. Interpretability refers to the extent to which a human can understand the cause of a decision made by a model, while explainability involves providing understandable justifications for the model's outputs. These concepts are essential for ensuring that AI systems are not only powerful but also trustworthy and fair (Kaur et al., 2022; Leblanc & Germain, 2023).In high-stakes areas such as healthcare, finance, and autonomous systems, the implications of using black-box models—AI systems whose decision-making processes are opaque—are particularly significant. Black-box models can achieve high levels of accuracy and performance; however, their lack of transparency poses ethical, legal, and practical challenges (Mesinovic et al., 2023; Nauta et al., 2022). The opacity of these models can lead to issues such as the propagation of biases, unfair outcomes, lack of accountability, and difficulties in regulatory compliance. Additionally, without clear explanations for their decisions, these models can erode trust among users and stakeholders, ultimately hindering the broader adoption of AI technologies (Kaur et al., 2022).Recent advancements in explainable AI (XAI) techniques have shown promise in addressing these challenges by enhancing model interpretability and transparency. Techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are increasingly being used to provide insights into model predictions, making it easier to understand and trust AI systems (Custode & Iacca, 2022; Swathi & Challa, 2023). These advancements are crucial for ensuring that AI systems are not only efficient but also fair and accountable in their decision-making processes.

To address these challenges, this paper will next discuss the ethical and legal implications of black-box models, particularly in high-stakes areas such as healthcare, finance, and autonomous systems. The inherent opacity of these models poses several critical issues, including the propagation of biases, unfair outcomes, lack of accountability, and difficulties in regulatory compliance.

We will explore how black-box models can amplify existing biases in data, leading to discriminatory outcomes in critical sectors. For instance, biased training data can result in less accurate healthcare diagnoses for certain demographic groups or unfair lending practices in finance. The opacity of these models complicates the identification and correction of such biases. Additionally, the paper will discuss the significant and sometimes detrimental effects of black-box model decisions on individuals' lives, such as unjust outcomes in the criminal justice system or critical errors in autonomous driving. These issues underscore the need for transparent decision-making processes to maintain fairness and public trust. The lack of accountability is another major concern, especially in regulated industries like finance and healthcare. The inability to trace decisions back to specific inputs and rules hinders effective accountability, which is crucial for ethical governance and regulatory compliance.

Finally, we will examine the evolving regulatory frameworks that emphasize the necessity for transparency in AI systems, such as the European Union's GDPR and the Fair Credit Reporting Act in the United States. These regulations demand the development of interpretable models that provide clear and understandable explanations for their decisions, ensuring legal compliance and fostering trust among users and stakeholders. Ensuring that AI systems are transparent and understandable is essential for mitigating biases, ensuring fairness, maintaining accountability, complying with regulations, building trust, and making ethical decisions. As AI continues to be integrated into critical areas of society, the need for interpretable and explainable models will only become more pronounced.

II. ETHICAL AND LEGAL IMPLICATIONS OF BLACK-BOX MODELS

Black-box models pose significant ethical and legal challenges, particularly in high-stakes areas such as healthcare, finance, and autonomous systems. The inherent opacity of these models means that their decision-making processes are not easily understood, which can lead to several critical issues (Maeda et al. (2021). One of the primary ethical concerns with black-box models is their potential to propagate and even amplify existing biases in the data. Since these models often learn from historical data, any biases present in the training data can be reflected in the model's predictions. In healthcare, for instance, if the training data contains biases against certain demographic groups, the model might provide less accurate diagnoses or treatment recommendations for those groups (Roa et al., 2023). Similarly, in finance, biased credit scoring models can lead to discriminatory lending practices (Ross & Shin, 2024).

The lack of interpretability makes it challenging to identify and correct these biases, leading to unfair outcomes. The decisions made by black-box models can have significant and sometimes detrimental effects on individuals' lives. For example, the inability to understand why a self-driving car made a particular decision can be a matter of life and death (Eberle et al. 2021). In the criminal justice system, predictive policing models or risk assessment tools that are not interpretable can result in unjust outcomes, such as wrongful arrests or unfair sentencing (Maeda et al., 2021). The opacity of these models can undermine the fairness of the decisions they support, leading to a loss of public trust. Accountability is another critical issue associated with black-box models. When decisions made by these models are not transparent, it becomes difficult to hold the appropriate parties responsible for the outcomes (Peters, 2023). In scenarios where a model's decision leads to an adverse event, the lack of clear explanations can prevent effective accountability. This is particularly problematic in regulated industries like finance and healthcare, where accountability is a legal requirement. Regulatory bodies are increasingly emphasizing the need for transparency in AI systems to ensure that decisions can be traced and justified (Wischmeyer & Rademacher, 2020).

Regulations in various sectors are evolving to address the transparency and accountability of AI systems. For example, the European Union's General Data Protection Regulation (GDPR) (2018) includes provisions for the "right to explanation," where individuals can request an explanation for decisions made by automated systems. In the financial sector, the Fair Credit Reporting Act (FCRA) in the United States requires that adverse actions based on credit scores be explained to consumers (Smith & Bartholomew, 2023). These regulatory frameworks necessitate the development of interpretable models that can provide clear and understandable explanations for their decisions.For AI systems to be widely adopted, especially in critical areas, they must be trusted by their users. Trust is built through transparency and understanding. When users and stakeholders can see and understand how a model makes decisions, they are more likely to trust and rely on the system (Vainio-Pekka et al.,, 2023). In healthcare, for example, doctors are more likely to use AI tools if they can understand how the tool arrived at a diagnosis or treatment recommendation (Beltramin et al., 2022). Similarly, in finance, consumers are more likely to trust and accept decisions made by automated systems if they can understand the reasoning behind those decisions (Messud & Chambefort, 2020).

Interpretability and explainability are crucial for ensuring that AI systems make ethical decisions. Ethical decision-making in AI involves not only making accurate predictions but also considering the broader impact of those decisions on individuals and society. Interpretable models allow for the incorporation of ethical considerations into the decision-making process (Amann et al., 2020). For instance, in loan approval processes, an interpretable model can help ensure that decisions are made fairly and equitably across different demographic groups (Mujtaba & Mahapatra, 2020). The need for interpretability and explainability in AI systems, particularly in ethical decision-making, is a recurring theme in the literature. Kumar and Sharma (2020) and Mcdermid et al. (2021) both emphasize the importance of transparency and accountability in these systems, with the latter linking these technical aspects to ethical considerations. Amann et al. (2020) and Dilip et al. (2022) further explore the role of explainability in healthcare and the incorporation of trustworthiness and ethics in AI systems, respectively. Vainio-Pekka et al. (2023) provides a systematic mapping of the research field of AI ethics, highlighting the need for a common framework. Meske et al. (2022) identifies opposing effects of AI explainability and proposes a framework for managing these tensions. Lastly, Lo Piano (2020) discusses the ethical implications of AI-driven decision-making, particularly in risk assessment and autonomous vehicles.

The ethical and legal implications of black-box models underscore the importance of interpretability and explainability in AI systems. Ensuring that these systems are transparent and understandable is essential for mitigating biases, ensuring fairness, maintaining accountability, complying with regulations, building trust, and making ethical decisions. As AI continues to be integrated into critical areas of society, the need for interpretable and explainable models will only become more pronounced.

III. NEED FOR TRANSPARENCY AND TRUST IN AI SYSTEMS

Transparency in AI systems is crucial for fostering trust among users and stakeholders. Trust is a foundational element for the widespread adoption and integration of AI technologies across various sectors. When users have a clear understanding of how a model arrives at its decisions, they are more likely to trust and accept its outputs. This section explores the multifaceted benefits of transparency in AI systems and the critical role it plays in building trust. Trust is a critical factor in the widespread adoption of AI technologies across various sectors (Gupta, 2023). Factors influencing trust include the difficulty of the task, perceived performance, and success/failure of the task (Konidena et al., 2024). Key requirements for trust in AI include access to knowledge, transparency, explainability, certification, and self-imposed standards and guidelines (Ferrara, 2023). User trust is influenced by socio-ethical considerations, technical and design features, and user characteristics. Trust in AI technologies significantly impacts their acceptance (Loni et al., 2020). Trust in AI is crucial in critical sectors such as healthcare, defense, and security (Rao et al., 2023). The concept of trustworthy AI is based on the principles of beneficence, non-maleficence, autonomy, justice, and explicability (Bianchini & Scarselli, 2014; Ferrara, 2023).

Transparency enhances user confidence in AI systems by demystifying the decision-making process. For example, in healthcare, tools like LIME (Local Interpretable Model-agnostic Explanations) allow doctors to see which features influenced a diagnosis, thereby increasing their trust in AI-based diagnostic tools. In finance, the use of SHAP (SHapley Additive exPlanations) in credit scoring models provides clear insights into the factors affecting loan approval decisions, reassuring applicants of the process's fairness. Visual tools like heatmaps and saliency maps further illustrate the model's focus areas, aiding in user understanding and confidence (Rao et al., 2023).

Transparency enables users to make informed decisions based on the model’s outputs. For instance, in healthcare, a transparent AI system that explains its diagnosis can help doctors decide on the best course of treatment. In finance, transparent credit scoring models allow loan officers to understand the factors influencing creditworthiness, enabling them to make more informed lending decisions. This informed decision-making is crucial for the effective use of AI technologies in high-stakes environments (Mouton & Davel, 2022).Transparent AI systems facilitate accountability by making it clear how decisions are made. This is essential for ethical responsibility, as it allows stakeholders to trace decisions back to specific inputs and rules within the model. In the event of an adverse outcome, transparency ensures that the appropriate parties can be held accountable, which is critical for ethical governance and compliance with regulatory standards. For example, in autonomous driving, understanding the decision-making process of a self-driving car can help determine liability in the case of an accident (Chaudhuri, 2019).As regulatory bodies increasingly emphasize the need for transparency in AI systems, transparency becomes essential for legal compliance. Regulations such as the European Union’s General Data Protection Regulation (GDPR) and the proposed AI Act mandate that AI systems provide explanations for their decisions.

These regulations aim to protect individuals' rights and ensure that AI systems operate fairly and transparently. Transparent models are therefore better positioned to comply with these regulatory requirements, reducing legal risks and fostering trust among stakeholders (Chaudhuri, 2019). Transparency in AI systems also helps identify and mitigate risks and biases. By understanding how a model processes data and arrives at its predictions, developers and users can detect potential biases in the training data or model structure. This is particularly important in applications such as hiring, where biased algorithms can perpetuate discrimination. Transparent models enable stakeholders to identify and address these issues, ensuring that AI systems operate fairly and equitably (Ferrara, 2023; Gupta, 2023).Public trust in AI systems is crucial for their acceptance and integration into society. Transparent AI systems that explain their decisions can alleviate public concerns about the potential negative impacts of AI, such as job displacement, privacy invasion, and loss of control (Chaudhuri, 2019). By demonstrating how AI systems make decisions and highlighting their benefits, transparency helps build public trust and acceptance, paving the way for the broader adoption of AI technologies. Transparency encourages the development of ethical AI systems by making it clear how decisions are made and highlighting any potential ethical concerns. Developers who prioritize transparency are more likely to consider the ethical implications of their models and take steps to ensure that their systems operate fairly and responsibly. This ethical focus not only builds trust among users but also promotes the development of AI technologies that are aligned with societal values and norms (Akinrinola et al., 2024; Olatoye et al., 2024).

Transparency is a key factor in the development of ethical AI systems, as it allows for the clear understanding of decision-making processes and the identification of potential ethical concerns (Konidena et al., 2024). However, the implementation of transparency in AI systems is complex, requiring a balance between information provision and contextual factors (Mouton & Davel, 2022). Despite this complexity, the value of transparency in AI is widely recognized, with calls for both "outward" and "functional" transparency (Gupta, 2023). The importance of transparency is further underscored by its inclusion in ethical principles of AI, alongside justice, fairness, responsibility, non-maleficence, and privacy.

IV. CASE STUDIES

Environmental monitoring through interpretable AI models offers valuable insights for policymaking and public awareness. A deep neural network was developed to predict air quality indices using data on weather conditions, pollutant levels, and traffic. By applying SHAP to analyze feature importance, the model identified weather patterns and traffic congestion as key factors impacting air quality predictions (Gebreyesus et al., 2023). The interpretability enabled policymakers to understand the drivers of poor air quality and develop targeted interventions. It also facilitated communicating the findings to the public, fostering trust and encouraging community participation in pollution reduction efforts. Interpretability techniques have also been applied to understand climate emulation models. Xu et al. (2021) employed post-hoc local explanation methods and feature importance on a deep learning emulator predicting sea surface temperature variations. The analysis revealed that the emulator's predictions were predominantly influenced by a localized region around the target area, with the important area extending further back in time for longer prediction lead times. Ablation experiments verified these findings, suggesting local processes dominate climate emulation with minimal influence from remote teleconnections.

In the education domain, a deep neural network for predicting student performance utilized LIME to provide insights into individual predictions based on factors like attendance, assignments, and online engagement (Kumar and Sharma, 2020). The interpretability allows educators to understand the influential factors for each student, enabling personalized teaching strategies and timely interventions for at-risk students. The transparency also helps gain trust from educators and parents in the AI system's recommendations.

V. IMPLICATIONS FOR PRACTITIONERS AND RESEARCHERS IN DESIGNING INTERPRETABLE DNNS

For practitioners and researchers aiming to design interpretable DNNs, it is crucial to adopt domain-specific interpretability tools tailored to the task at hand. For instance, in environmental monitoring, tools like SHAP can provide clear attribution of factors affecting air quality predictions, while in education, LIME offers insights into the factors influencing student performance assessments. Employing these targeted tools enhances the relevance and accuracy of model explanations, fostering trust and compliance from stakeholders.To improve transparency, practitioners should incorporate attention mechanisms into their models, especially for tasks like energy management where highlighting the relevant features influencing energy distribution decisions can help users understand and trust the model's reasoning. However, it is important to strike a balance between model complexity and interpretability. Implementing hybrid models that combine shallow interpretable layers with

deeper complex structures can achieve high performance without sacrificing transparency. Visualization techniques like heatmaps are invaluable for identifying important features and debugging models, aiding in understanding model behavior and facilitating validation and refinement. Continuous evaluation and iterative improvement of model explanations based on user feedback and new data is also essential to ensure models remain accurate, reliable, and interpretable over time.

❖ **Adopt Domain-Specific Interpretability Tools:**
✓ **Action:** Select for tools like SHAP environmental monitoring to provide clear attribution of factors affecting air quality, or LIME for education to offer insights into student performance.
✓ **Benefit:** Tailored tools enhance the relevance and accuracy of model explanations, fostering trust and compliance.

❖ **Incorporate Attention Mechanisms:**
✓ **Action:** Use attention mechanisms in energy management tasks to improve transparency by highlighting relevant features influencing energy distribution.
✓ **Benefit:** Improved model transparency helps users understand and trust the model's decision-making process.

❖ **Balance Model Complexity and Interpretability:**
✓ **Action:** Implement hybrid models combining shallow interpretable layers with deeper complex structures.
✓ **Benefit:** Achieve high performance without sacrificing interpretability, ensuring models are both effective and transparent.
✓
❖ **Utilize Visualization Techniques:**
✓ **Action:** Employ visualization tools like heatmaps to identify important features and debug models.
✓ **Benefit:** Visualization aids in understanding model behavior, facilitating validation and refinement.

❖ **Continuous Evaluation and Refinement:**
✓ **Action:** Regularly test and improve model explanations based on user feedback and new data.
✓ **Benefit:** Ongoing refinement ensures models remain accurate, reliable, and interpretable.

Looking ahead, several recommendations can help balance performance and interpretability in real-world applications. Hybrid models that leverage the strengths of both shallow and deep architectures should be implemented to combine high performance with a degree of transparency. Visualization tools should be an integral part of the model development and validation process, aiding in understanding and debugging. Continuous evaluation and refinement of models based on interpretability assessments and feedback from domain experts and end-users is crucial. Fostering interdisciplinary collaboration among computer scientists, domain experts, ethicists, and others is vital to ensure models are technically sound, practically relevant, and ethically aligned. Moreover, there is a pressing need for standardized, multi-dimensional benchmarks to evaluate interpretability across different contexts. These benchmarks should be flexible, adaptable, and provide a comprehensive framework for assessing model transparency and performance.

❖ **Implement Hybrid Models:**
✓ **Recommendation:** Combine shallow interpretable layers with deeper, more complex structures to leverage the strengths of both approaches, ensuring high performance while maintaining a degree of transparency.
❖ **Utilize Visualization Techniques:**
✓ **Recommendation:** Employ visualization tools like heatmaps and feature maps to aid in understanding model behavior and debugging. These tools should be an integral part of the model development and validation process.
❖ **Continuous Evaluation and Refinement:**
✓ **Recommendation:** Models should be subject to ongoing evaluation and refinement based on interpretability assessments. This involves iteratively testing and improving the model's explanations, incorporating feedback from domain experts and end-users.
❖ **Foster Interdisciplinary Collaboration:**
✓ **Recommendation:** Developing interpretable DNNs requires collaboration among computer scientists, domain experts, and ethicists. This interdisciplinary approach ensures that models are not only technically sound but also practically relevant and ethically aligned with real-world applications.

❖ **Develop Standardized Benchmarks:**

✓ **Recommendation:** There is a need for standardized, multi-dimensional benchmarks to evaluate interpretability across different contexts. These benchmarks should be flexible and adaptable to various domains, providing a comprehensive framework for assessing model transparency and performance.

Future research should focus on developing advanced visualization tools that can demystify complex models and provide detailed, real-time insights into decision-making processes. Exploring hybrid model architectures that combine interpretable and complex layers is another promising avenue. Interdisciplinary research collaborations should be encouraged to develop holistic interpretability solutions that are not only technically robust but also ethically sound and practically relevant. Standardized and multi-dimensional benchmarks for evaluating interpretability across domains are also needed. Finally, researchers should investigate methods for integrating interpretability considerations throughout the entire model development lifecycle, from design and training to evaluation, ensuring transparency is a core objective.

❖ **Development of Advanced Visualization Tools:**

✓ **Focus:** Create new visualization techniques that can demystify complex models. These tools should provide detailed, real-time insights into model behavior and decision-making processes, making them accessible and useful for both researchers and practitioners.

❖ **Exploration of Hybrid Model Architectures:**

✓ **Focus:** Investigate hybrid architectures that combine shallow, interpretable layers with deeper, more complex ones to offer a promising path forward. Such models can achieve high performance while retaining a degree of transparency, making them suitable for a wide range of applications.

❖ **Interdisciplinary Research Collaborations:**

✓ **Focus:** Encourage collaborations between computer scientists, domain experts, ethicists, and psychologists to develop more holistic interpretability solutions. These collaborations can ensure that models are not only technically robust but also ethically sound and practically relevant.

❖ **Standardized and Multi-Dimensional Benchmarks:**

✓ **Focus:** There is a pressing need for the creation of standardized benchmarks that evaluate interpretability across multiple dimensions. These benchmarks should be adaptable to various domains and should consider factors such as clarity, relevance, consistency, and actionability. Developing such benchmarks will provide a clearer framework for assessing and improving the interpretability of DNNs.

❖ **Integration of Interpretability in Model Development Lifecycle:**

✓ **Focus:** Explore methods for integrating interpretability considerations throughout the entire model development lifecycle. This includes incorporating interpretability metrics during the design, training, and evaluation phases, ensuring that models are developed with transparency as a core objective.

By focusing on these future directions, researchers and practitioners can advance the development of interpretable and transparent DNNs, ensuring that AI systems are not only powerful but also trustworthy and ethically aligned with societal needs.

VI. CONCLUSION

The interpretability and explainability of deep neural networks (DNNs) are paramount as AI becomes increasingly integrated into critical domains like healthcare, finance, and autonomous systems. The inherent opacity of black-box models can lead to serious ethical issues like propagating biases, generating unfair outcomes, lacking accountability, and hindering regulatory compliance. Ensuring AI systems are transparent and their decision-making processes are interpretable is crucial for fostering trust, making ethical decisions, and encouraging wider adoption of these powerful technologies.This study explored different architectural components and techniques that can enhance the interpretability of DNNs. Case studies demonstrated the practical value of interpretable models, such as using SHAP to understand key drivers of air quality for policymaking or applying LIME to provide personalized insights on student performance. Recommendations for practitioners emphasized adopting domain-specific interpretability tools, incorporating attention mechanisms, balancing model complexity with transparency through hybrid architectures, utilizing visualizations, and continuously refining models based on user feedback.

Looking ahead, future research should focus on developing advanced visualization techniques that demystify complex models, exploring hybrid architectures that combine interpretable and complex layers, fostering interdisciplinary collaborations to create holistic solutions, establishing standardized benchmarks to evaluate interpretability across domains, and integrating interpretability considerations throughout the entire model development lifecycle. By prioritizing interpretability and transparency, researchers and developers can create AI systems that are not only accurate and high-performing but also trustworthy, ethical, and aligned with societal values. Navigating the complexities of DNNs while ensuring transparency is crucial for realizing the immense potential of AI in transforming critical sectors for the betterment of humanity.

REFERENCES

1. Akinrinola, O., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews, 18*(3), 050-058.
2. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making, 20*, 1-9.
3. Beltramin, D., Lamas, E., & Bousquet, C. (2022). Ethical issues in the utilization of black boxes for artificial intelligence in medicine. In *Advances in Informatics, Management and Technology in Healthcare* (pp. 249-252). IOS Press.
4. Bianchini, M., & Scarselli, F. (2014). On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems, 25*(8), 1553-1565.
5. Chaudhuri, A. (2019). Some insights and observations on depth issues in deep learning networks. In *Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I* 15 (pp. 583-595). Springer International Publishing. https://doi.org/10.1007/978-3-030-20521-8_48
6. Custode, L. L., & Iacca, G. (2022, July). Interpretable pipelines with evolutionary optimized modules for reinforcement learning tasks with visual inputs. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 224-227).
7. Dilip, G., Guttula, R., Rajeyyagari, S., Pandey, R. R., Bora, A., R Kshirsagar, P., ... & Sundramurthy, V. P. (2022). Artificial Intelligence-Based Smart Comrade Robot for Elders Healthcare with Strait Rescue System. *Journal of Healthcare Engineering, 2022*(1), 9904870.
8. Dong, Y., Su, H., Zhu, J., & Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 975-983. https://doi.org/10.1109/CVPR.2017.110
9. Eberle, H., Zhang, B., Teodorescu, C., Walker, G., & Carlson, T. (2021). An 'ethical black box' learning from disagreement in shared control systems. *IEEE International Conference on Systems, Man, and Cybernetics (SMC).* 10.1109/SMC52423.2021.9658964
10. Farahani, F. V., Fiok, K., Lahijanian, B., Karwowski, W., & Douglas, P. K. (2022). Explainable AI: A review of applications to neuroimaging data. *Frontiers in Neuroscience*, 16. https://doi.org/10.3389/fnins.2022.906290
11. Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *ArXiv, abs/2304.07683*. https://doi.org/10.3390/sci6010003
12. GDPR, G. D. P. R. (2018). General data protection regulation. URL: *https://gdpr-info. eu/[accessed 2020-11-21]*.
13. Gebreyesus, Y., Dalton, D., De Chiara, D., Chinnici, M., & Chinnici, A. (2024). AI for Automating Data Center Operations: Model Explainability in the Data Centre Context Using Shapley Additive Explanations (SHAP). *Electronics, 13*(9), 1628.
14. Gupta, N. (2023). Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications. *Revista Review Index Journal of Multidisciplinary, 3*(2), 24-35. https://doi.org/10.31305/rrijm2023.v03.n02.004
15. Kaur, H., Adar, E., Gilbert, E., & Lampe, C. (2022). Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3531146.3533135
16. Konidena, B. K., Malaiyappan, J. N. A., & Tadimarri, A. (2024). Ethical Considerations in the Development and Deployment of AI Systems. *European Journal of Technology, 8*(2), 41-53. https://doi.org/10.47672/ejt.1890

17. Kumar, P., & Sharma, M. (2020). Predicting Academic performance of international students using machine learning techniques and human interpretable explanations using LIME—Case study of an Indian University. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2019*, Volume 1 (pp. 289-303). Springer Singapore. https://doi.org/10.1007/978-981-15-1286-5_25
18. Leblanc, B., & Germain, P. (2023). Interpretability in Machine Learning: on the Interplay with Explainability, Predictive Performances and Models. *ArXiv, abs/2311.11491*. https://doi.org/10.48550/arXiv.2311.11491
19. Liang, Y., Li, S., Yan, C., Li, M., & Jiang, C. (2021). Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing, 419,* 168-182. https://doi.org/10.1016/j.neucom.2020.08.011
20. Loni, M., Sinaei, S., Zoljodi, A., Daneshtalab, M., & Sjödin, M. (2020). DeepMaker: A multi-objective optimization framework for deep neural networks in embedded systems. Microprocess. *Microsystems, 73*, 102989. https://doi.org/10.1016/j.micpro.2020.102989
21. Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications, 7*(1), 1-7.
22. Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc. Red Hook, NY, United States.
23. Maeda, E. E., Haapasaari, P., Helle, I., Lehikoinen, A., Voinov, A., & Kuikka, S. (2021). Black boxes and the role of modeling in environmental policy making. *Frontiers in Environmental Science, 9*, 629336. https://doi.org/10.3389/fenvs.2021.629336
24. McDermid, J. A., Jia, Y., Porter, Z., & Habli, I. (2021). Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A, 379*(2207), 20200363.
25. Mesinovic, M., Watkinson, P., & Zhu, T. (2023). Explainable AI for clinical risk prediction: a survey of concepts, methods, and modalities. *ArXiv, abs/2308.08407*. https://doi.org/10.48550/arXiv.2308.08407
26. Meske, C., Abedin, B., Klier, M., & Rabhi, F. (2022). Explainable and responsible artificial intelligence. *Electronic Markets, 32*(4), 2103-2106.
27. Messud, J., & Chambefort, M. (2020, November). Understanding how a deep neural network architecture choice can be related to a seismic processing task. In *First EAGE Digitalization Conference and Exhibition* (Vol. 2020, No. 1, pp. 1-5). European Association of Geoscientists & Engineers.
28. Mouton, C., & Davel, M. H. (2021, December). Exploring layerwise decision making in DNNs. In *Southern African Conference for Artificial Intelligence Research* (pp. 140-155). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-95070-5_10
29. Mujtaba, D. F., & Mahapatra, N. R. (2020, December). Artificial intelligence in computerized adaptive testing. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 649-654). IEEE.
30. Olatoye, F. O., Awonuga, K. F., Mhlongo, N. Z., Ibeh, C. V., Elufioye, O. A., & Ndubuisi, N. L. (2024). AI and ethics in business: A comprehensive review of responsible AI practices and corporate responsibility. *International Journal of Science and Research Archive, 11*(1), 1433-1443.
31. Peters, U. (2023). Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque. *AI and Ethics, 3*(3), 963-974.
32. Rao, T., Agarwal, S., & Singh, N. (2023). An empirical evaluation of Shapley additive explanations: A military implication. *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 10, 1390-1397. https://doi.org/10.1109/UPCON59197.2023.10434608
33. Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Association for Computing Machinery.* 1135–1144. https://doi.org/10.1145/2939672.2939778
34. Ross, D. G., & Shin, D. H. (2024). Do financial market frictions hurt the performance of women-led ventures? A meta-analytic investigation. *Strategic Management Journal, 45*(3), 507-534.
35. Sheu, Y. (2020). Illuminating the black box: Interpreting deep neural network models for psychiatric research. *Frontiers in Psychiatry*, 11. https://doi.org/10.3389/fpsyt.2020.551299
36. Smith, A. M., & Bartholomew, L. C. (2023). Fair Credit Reporting Act Update-2022. *Bus. Law.*, 78, 539.

37. Swathi, Y., & Challa, M. (2023, June). A Comparative Analysis of Explainable AI Techniques for Enhanced Model Interpretability. In *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)* (pp. 229-234). IEEE.
38. Vainio-Pekka, H., Agbese, M. O. O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P. (2023). The role of explainable ai in the research field of ai ethics. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 1-39.
39. Wischmeyer, T., & Rademacher, T. (Eds.). (2020). *Regulating artificial intelligence* (Vol. 1, No. 1, pp. 307-321). Cham: Springer.
40. Xu, W., Luo, X., Ren, Y., Park, J. H., & Yoo, S. (2021). Feature importance in a deep learning climate emulator. In *AI: Modeling Oceans and Climate Change Workshop at ICLR 2021*. Computational Science Initiative.
41. Zhang, W., Zhang, L., Zhang, Z., & Sun, M. (2021). IBD: The metrics and evaluation method for DNN processor benchmark while doing inference task. *J. Intell. Fuzzy Syst., 40*, 9949-9961. https://doi.org/10.3233/JIFS-202552