iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** X **Month of publication:** October 2024

DOI: https://doi.org/10.22214/ijraset.2024.64567

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue X Oct 2024- Available at www.ijraset.com

Advancing Trustworthy AI: Leveraging Explainable AI (XAI) to Drive Safer, Fairer, and More Reliable Intelligent Systems for Next-Generation Technological Innovation

Aadhith Rajinikanth[1], Akash Kalita[2]

I. INTRODUCTION

The rapid growth of artificial intelligence (AI) across multiple sectors, including healthcare and transportation, has brought significant advancements but also challenges related to safety, reliability, fairness, and trustworthiness. Many AI models, particularly deep neural networks, operate as "black boxes" with mechanisms that are not transparent, making it hard to understand their decision-making processes and potential biases. This opacity can lead to unintended outcomes and diminish public trust in AI technologies.

To address these concerns, the field of Explainable AI (XAI) has emerged, focusing on making AI systems more interpretable and transparent. XAI aims to clarify the logic behind AI decisions, thereby improving safety, fairness, and accountability in AI applications. This increased transparency not only builds trust but also promotes the responsible use of AI technologies.

We chose to focus on this topic because, as AI becomes more integrated into various aspects of our lives, it is imperative that these systems are developed and utilized in a responsible and ethical manner. XAI plays a crucial role by enabling stakeholders to understand, review, and regulate AI decision-making processes, ensuring that they align with human values and societal standards, and helping to prevent biases and errors.

II. AI SAFETY AND EXPLAINABLE AI(XAI): FOUNDATIONS

A. Definitions and Objectives of AI Safety

AI safety is a crucial field that ensures artificial intelligence systems are developed and operated in ways that align with human values, ethics, and societal standards, while minimizing risks and unintended consequences (Amodei et al., 2016). The main goals of AI safety include ensuring the reliability and robustness of AI systems to avoid failures or errors that could cause harm. It also aims to enhance transparency and interpretability, making it easier for humans to understand how AI models make decisions, which is essential for building trust and accountability (Hendrycks et al., 2021). Additionally, AI safety addresses fairness and bias, striving to ensure AI systems treat all individuals fairly, regardless of race, gender, or other characteristics (Mehrabi et al., 2021). Moreover, AI safety focuses on protecting privacy and securing sensitive data used by AI systems, and it emphasizes accountability, setting clear responsibilities for the actions and decisions of AI systems (Amodei et al., 2016).

B. Overview of XAI Approaches and Interpretability Concepts

Explainable AI (XAI) is dedicated to making AI systems more understandable and transparent, enabling humans to grasp the logic behind their decisions and outputs (Gilpin et al., 2018). XAI methods are divided into two main types: inherently interpretable models and post-hoc explanation techniques. Inherently interpretable models, such as decision trees, rule-based systems, and linear models, are straightforward in their structure and decision-making processes, making them easier to understand (Molnar, 2022). In contrast, post-hoc explanation techniques provide insights into more complex, "black-box" models like deep neural networks. These include methods such as feature attribution, saliency maps, and counterfactual explanations that clarify how decisions are made (Ribeiro et al., 2016). XAI also involves different interpretability concepts, such as local and global explanations, model-agnostic and model-specific techniques, and various metrics to evaluate the quality and accuracy of explanations (Molnar, 2022). Local explanations focus on individual predictions, while global explanations address the model's overall behavior. Model-agnostic methods can be used with any model type, whereas model-specific techniques are designed for particular architectures. Evaluation metrics assess explanations on their fidelity, consistency, and how understandable they are to humans (Doshi-Velez & Kim, 2017).

©IJRASET: All Rights are Reserved

C. Synergies between XAI and AI Safety, Fairness, and Trust Goals

XAI is key to achieving AI safety, fairness, and trust (Arrieta et al., 2020). By making the decision-making processes of AI systems clearer, XAI enhances safety and reliability by helping humans understand, verify, and correct AI behavior. It also promotes fairness by exposing and addressing bias, aiding in the development of more equitable AI systems (Doshi-Velez & Kim, 2017). Moreover, XAI boosts trust and accountability by increasing transparency, allowing stakeholders to evaluate AI decisions against ethical standards and regulations. XAI also improves human-AI collaboration by making AI systems more approachable and comprehensible, fostering effective communication and cooperation. By providing explanations understandable to humans, XAI bridges the gap between complex AI models and human decision-making, supporting better integration of AI across different sectors (Doshi-Velez & Kim, 2017).

III. APPLYING XAI FOR SAFER AND MORE TRUSTWORTHY AI

A. Interpretability for Model Debugging and Error Analysis

Interpretability is a fundamental aspect of Explainable AI (XAI) that greatly contributes to model debugging and error analysis, thereby enhancing the development of safer and more trustworthy AI systems. By providing insights into the decision-making processes of AI models, interpretability techniques enable the identification and diagnosis of potential errors, biases, or anomalies within the model's behavior (Gilpin et al., 2018).

A crucial application of interpretability is model debugging, where developers utilize XAI techniques to comprehend why a model is making certain predictions or decisions. This can involve techniques such as feature attribution, which identifies the most influential features contributing to a particular output, or saliency maps, which visually highlight the areas of input data most relevant to the model's decision-making (Ribeiro et al., 2016). By examining these explanations, developers can pinpoint issues such as misjudged feature importance or unexpected patterns in the data, and take corrective actions to enhance the model's performance and reliability.

Error analysis is another critical application of interpretability within XAI. By utilizing the explanations provided by XAI techniques, developers can identify and investigate instances where the model makes incorrect predictions or decisions. This process may involve examining the feature attributions or saliency maps for specific examples to understand why the model failed, or analyzing patterns across multiple errors to identify systematic issues or weaknesses in the model's architecture or training data. Additionally, interpretability aids in the detection and mitigation of adversarial attacks or data poisoning, where malicious inputs are designed to mislead or manipulate the AI system. By examining the explanations for suspicious or anomalous predictions, developers can identify potential adversarial examples and implement appropriate countermeasures to enhance the robustness and security of their AI systems (Arrieta et al., 2020).

B. Detecting Bias, Fairness Issues Through Explanations

XAI plays a crucial role in detecting and mitigating bias and fairness issues in AI systems, which is essential for developing trustworthy and ethical AI solutions. By providing explanations for the decisions and predictions made by AI models, XAI techniques can reveal potential biases or discriminatory patterns that may be present in the data, algorithms, or model outputs (Mehrabi et al., 2021). One approach to detecting bias through explanations involves analyzing the feature attributions or saliency maps for protected attributes, such as race, gender, or age. If these attributes significantly influence the model's decisions, it may indicate the presence of bias or unfair treatment toward certain groups.

Additionally, by examining explanations across different subgroups or demographics, researchers can identify disparities in the model's performance or decision-making processes, which could point to potential fairness issues. Another technique involves the use of counterfactual explanations, which provide insights into how the model's output would change if certain input features were modified (Wachter et al., 2018). By generating counterfactual explanations for individuals from different protected groups, researchers can assess whether the model's decisions are consistent and fair across these groups, or if there are disparities in the required changes or conditions for achieving a desired outcome.

Furthermore, XAI techniques can be used to audit and evaluate the fairness of AI systems by comparing the explanations and decision-making processes against established fairness metrics and criteria. This process can involve assessing individual fairness, which ensures that similar individuals are treated similarly, or group fairness, which focuses on ensuring statistical parity or equal opportunity across different demographic groups (Arrieta et al., 2020).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Using XAI to Enhance Robustness, Reliability, and Accountability

XAI is integral to enhancing the robustness, reliability, and accountability of AI systems, which are essential components of trustworthy and responsible AI. By providing explanations and insights into the decision-making processes of AI models, XAI techniques can help identify potential vulnerabilities, errors, or weaknesses, enabling developers to take corrective actions and improve the overall reliability and robustness of their systems. One way in which XAI contributes to robustness is by enabling the detection and mitigation of adversarial attacks or data poisoning attempts. By examining the explanations for suspicious or anomalous predictions, developers can identify potential adversarial examples and take appropriate countermeasures, such as data sanitization or model hardening, to enhance the security and resilience of their AI systems.

XAI techniques can also be used to assess the reliability and consistency of AI models across different input domains or scenarios. By analyzing the explanations and decision-making processes for a diverse range of inputs, developers can identify potential edge cases or corner cases where the model may perform poorly or produce unreliable outputs. This information can then be used to refine and improve the model's training data, architecture, or decision boundaries, ultimately enhancing its overall reliability and performance. Furthermore, XAI plays a crucial role in fostering accountability and transparency in AI systems, which are essential for building trust and ensuring responsible deployment. By providing explanations for the decisions and predictions made by AI models, XAI techniques enable stakeholders, regulators, and end-users to scrutinize and understand the reasoning behind these outputs. This transparency can help identify potential issues, biases, or ethical concerns, and facilitate the implementation of appropriate governance frameworks and accountability measures.

IV. SPECIFIC XAI MODELS AND TECHNIQUES

A. Inherently Interpretable Models (rule-based, decision trees, etc.)

Inherently interpretable models, such as rule-based and decision tree models, are designed to be transparent and easily understandable by humans. These models provide explanations through their inherent structure, making the decision-making process explicit and interpretable (Molnar, 2022). Rule-based models, like decision lists or rule sets, represent knowledge in the form of human-readable rules, which can be directly inspected and understood (Guidotti et al., 2018). Decision trees, on the other hand, recursively partition the input space into regions, with each leaf node representing a decision or prediction (Molnar, 2022). The hierarchical structure of decision trees allows for tracing the decision path, providing a clear explanation for each prediction.

B. Model-agnostic Explanation Methods for Black-box Models

Model-agnostic explanation methods aim to provide interpretability for black-box models, which are inherently opaque and difficult to understand. These methods treat the model as a black box and analyze its inputs and outputs to generate explanations (Molnar, 2022). Popular model-agnostic techniques include LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), which approximates the model's behavior locally with an interpretable model, and SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), which attributes the model's output to its input features based on game-theoretic concepts. These methods provide explanations that are model-agnostic and can be applied to any black-box model.

C. Evaluation Metrics for Interpretability and Explanation Quality

Evaluating the quality and effectiveness of interpretability and explanations is crucial for ensuring their usefulness and reliability. Various evaluation metrics have been proposed to assess different aspects of interpretability and explanation quality (Doshi-Velez & Kim, 2017). Some metrics focus on the fidelity of the explanations, measuring how accurately they represent the model's behavior (Molnar, 2022). Others assess the interpretability or comprehensibility of the explanations for humans, often through user studies or proxy measures (Doshi-Velez & Kim, 2017). Additionally, metrics like robustness and consistency evaluate the stability and coherence of explanations across different inputs or perturbations (Alvarez-Melis & Jaakkola, 2018).

V. FAIRNESS, ACCOUNTABILITY, AND BIAS MITIGATION WITH XAI

A. Using Explanations to audit for bias and Discrimination

Explanations from XAI techniques can be leveraged to audit machine learning models for potential biases and discriminatory behavior. By examining the feature importance and contributions towards a model's predictions, one can identify if certain sensitive attributes (e.g., race, gender) are unduly influencing the decisions (Doshi-Velez & Kim, 2017). This auditing process involves systematically probing the model with different input scenarios and analyzing the explanations to detect any unfair patterns or disparities across protected groups (Liao et al., 2020).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Additionally, counterfactual explanations can reveal the minimal changes required to receive a desired outcome, exposing potential discrimination if the changes disproportionately affect certain groups (Wachter et al., 2018).

XAI techniques also enable auditing for intersectional biases, where multiple sensitive attributes interact and compound discrimination (Buolamwini & Gebru, 2018). By examining the explanations for different combinations of sensitive attributes, one can identify if the model exhibits biases against specific intersectional groups (Kearns et al., 2019). This auditing process is crucial for ensuring fairness and non-discrimination, especially in high-stakes decision-making scenarios.

B. Counterfactual Reasoning For Recourse And Individual Fairness

Counterfactual explanations, which provide actionable suggestions for achieving a desired outcome, play a vital role in promoting individual fairness and recourse in machine learning systems (Ustun et al., 2019). These explanations inform individuals about the specific changes they can make to their input features to receive a favorable decision, empowering them with agency and transparency (Wachter et al., 2018).

Counterfactual reasoning techniques, such as CERTIFY (El Arras et al., 2022) and REVISE (Laugel et al., 2019), generate counterfactual explanations that satisfy various fairness constraints, ensuring that the suggested changes do not perpetuate discrimination or unfairness towards protected groups. By providing recourse opportunities tailored to individual circumstances while adhering to fairness principles, these techniques promote individual fairness and mitigate potential biases (Barocas et al., 2020).

C. Accountability, transparency, and ethical AI principles with XAI

XAI plays a crucial role in promoting accountability, transparency, and ethical AI principles by providing interpretable and understandable explanations for machine learning models' decisions (Arrieta et al., 2020). Explanations enable stakeholders, including developers, users, and affected individuals, to scrutinize the model's reasoning process, identify potential biases or unintended consequences, and hold the system accountable for its decisions (Doshi-Velez & Kim, 2017).

Furthermore, XAI aligns with ethical AI principles by promoting transparency, fairness, and non-discrimination (Robbins, 2019). By auditing models for biases and providing recourse opportunities through counterfactual explanations, XAI techniques help mitigate unfair or discriminatory behavior, fostering trust and ethical decision-making (Liao et al., 2020). Additionally, XAI facilitates human oversight and control over AI systems, enabling stakeholders to understand and validate the model's decisions, ensuring alignment with ethical principles and societal values (Arrieta et al., 2020).

VI. INTEGRATION OF XAI WITH OTHER AI AREAS

A. Combining XAI with causality for robust decision-making

Integrating XAI with causal reasoning techniques can lead to more robust and reliable decision-making in AI systems. Causal models aim to uncover the underlying causal relationships between variables, enabling a deeper understanding of the data-generating process and the ability to reason about interventions and counterfactuals (Pearl, 2009). By combining causal models with XAI methods, it becomes possible to provide explanations that go beyond mere correlations and capture the causal mechanisms driving the model's predictions (Karimi et al., 2020). One approach is to use causal models to generate counterfactual explanations, which explain how the model's output would change if certain input features were modified (Mahajan et al., 2019). These explanations can reveal the causal factors influencing the model's decisions, enabling users to understand the underlying reasoning and make more informed decisions. Additionally, causal models can be used to audit the model's behavior for potential biases or discriminatory patterns, by analyzing the causal pathways and identifying any unjustified dependencies on sensitive attributes (Chiappa, 2019).

B. Safe Exploration In Reinforcement Learning Via Explanations

In reinforcement learning (RL), safe exploration is crucial for agents to learn effective policies while avoiding potentially harmful or catastrophic actions. Explanations can play a vital role in enabling safe exploration by providing insights into the agent's decision-making process and identifying potential risks or uncertainties (Puiutta & Veith, 2020). One approach is to use explainable RL methods that generate human-interpretable explanations for the agent's actions and state-value estimates (Huang et al., 2019). These explanations can be used to monitor the agent's behavior and detect any anomalies or unexpected actions, allowing for human intervention or adjustment of the exploration strategy. Additionally, explanations can help identify areas of high uncertainty or risk, enabling the agent to focus its exploration efforts on safer regions of the state-action space (Srinivasan et al., 2020).

C. Explanations for human-AI collaboration and interaction

In scenarios where humans and AI systems collaborate or interact, explanations play a crucial role in fostering trust, transparency, and effective communication. By providing explanations for the AI system's decisions and recommendations, humans can better understand the reasoning behind the system's actions and make more informed decisions (Miller, 2019).

One approach is to develop interactive explanation interfaces that allow users to query the AI system for explanations and receive human-interpretable justifications (Liao et al., 2020). These interfaces can also enable users to provide feedback or corrections, which can be used to refine the AI system's decision-making process. Additionally, explanations can facilitate shared mental models between humans and AI systems, enabling more effective collaboration and coordination (Kulesza et al., 2013).

VII. FUTURE DIRECTIONS AND OPEN CHALLENGES

A. Scalability of XAI methods to complex models/data

As machine learning models become increasingly complex, with larger architectures and high-dimensional data, the scalability of XAI methods remains a significant challenge. Many existing XAI techniques struggle to provide meaningful explanations for deep neural networks with millions of parameters or high-dimensional data like images or text (Arrieta et al., 2020).

One promising direction is the development of more efficient and scalable XAI algorithms that can handle large-scale models and data. This may involve techniques like model distillation (Tan et al., 2018), which compresses complex models into simpler, interpretable forms, or hierarchical explanations that provide multi-level insights (Sundararajan et al., 2017). Additionally, leveraging advances in hardware acceleration and distributed computing could enable more efficient computation of explanations for complex models.

B. Objective Evaluation of Explanations and Human trust.

Evaluating the quality and effectiveness of explanations generated by XAI methods is a crucial challenge. While various metrics have been proposed, such as fidelity, consistency, and human-grounded evaluation (Doshi-Velez & Kim, 2017), there is a lack of standardized and objective evaluation frameworks. One promising direction is the development of comprehensive evaluation frameworks that incorporate multiple aspects of explanation quality, including fidelity to the model, human interpretability, and the ability to foster trust and understanding (Mohseni et al., 2020). These frameworks could leverage a combination of quantitative metrics, human subject studies, and task-based evaluations to provide a holistic assessment of explanations. Additionally, research into objective measures of human trust and understanding in response to explanations could inform the design of more effective XAI methods.

C. Emerging Frameworks for Responsible, Trustworthy AI with XAI

As AI systems become more prevalent in high-stakes decision-making scenarios, there is a growing need for frameworks that ensure responsible and trustworthy AI development and deployment. XAI plays a crucial role in enabling transparency, accountability, and ethical decision-making in AI systems (Arrieta et al., 2020). Emerging frameworks like the EU AI Act and the IEEE Ethically Aligned Design emphasize the importance of explainability, fairness, and robustness in AI systems (European Commission, 2021; IEEE, 2019). These frameworks provide guidelines and best practices for incorporating XAI techniques into the AI development lifecycle, enabling auditing for biases, providing recourse and explanations to affected individuals, and fostering trust and accountability. Additionally, research into integrating XAI with other AI areas, such as causality, safe exploration in reinforcement learning, and human-AI collaboration, could further enhance the development of responsible and trustworthy AI systems (Pearl, 2009; Puiutta & Veith, 2020; Miller, 2019).

VIII. CONCLUSION

Explainable AI (XAI) has emerged as a crucial component in the development and deployment of trustworthy and ethical AI systems. As AI models become increasingly complex and opaque, XAI techniques provide much-needed transparency, enabling stakeholders to understand the reasoning behind AI decisions and ensuring accountability.

XAI plays a vital role in enhancing AI safety, reliability, and fairness. By providing interpretable explanations, XAI methods enable model debugging, error analysis, and the detection of biases or unfair behavior. This capability is essential for building robust and reliable AI systems that can be trusted in high-stakes decision-making scenarios.

Moreover, XAI facilitates compliance with legal and ethical standards, promoting accountability and transparency. Explanations empower individuals affected by AI decisions, enabling them to exercise their right to explanation and ensuring fair treatment.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The integration of XAI with other AI areas, such as causality, reinforcement learning, and human-AI collaboration, further expands its potential. Causal explanations can uncover the underlying causal mechanisms driving AI decisions, while explanations in reinforcement learning enable safe exploration and risk mitigation. In human-AI collaboration, explanations foster trust, shared mental models, and effective communication.

However, challenges remain, including the scalability of XAI methods to complex models and high-dimensional data, the objective evaluation of explanation quality and human trust, and the development of standardized frameworks for responsible and trustworthy AI with XAI.

As AI systems continue to permeate various domains, the role of XAI in ensuring trustworthy and ethical AI will become increasingly crucial. By providing transparency, accountability, and alignment with human values, XAI represents a key enabler for the responsible development and deployment of AI technologies.

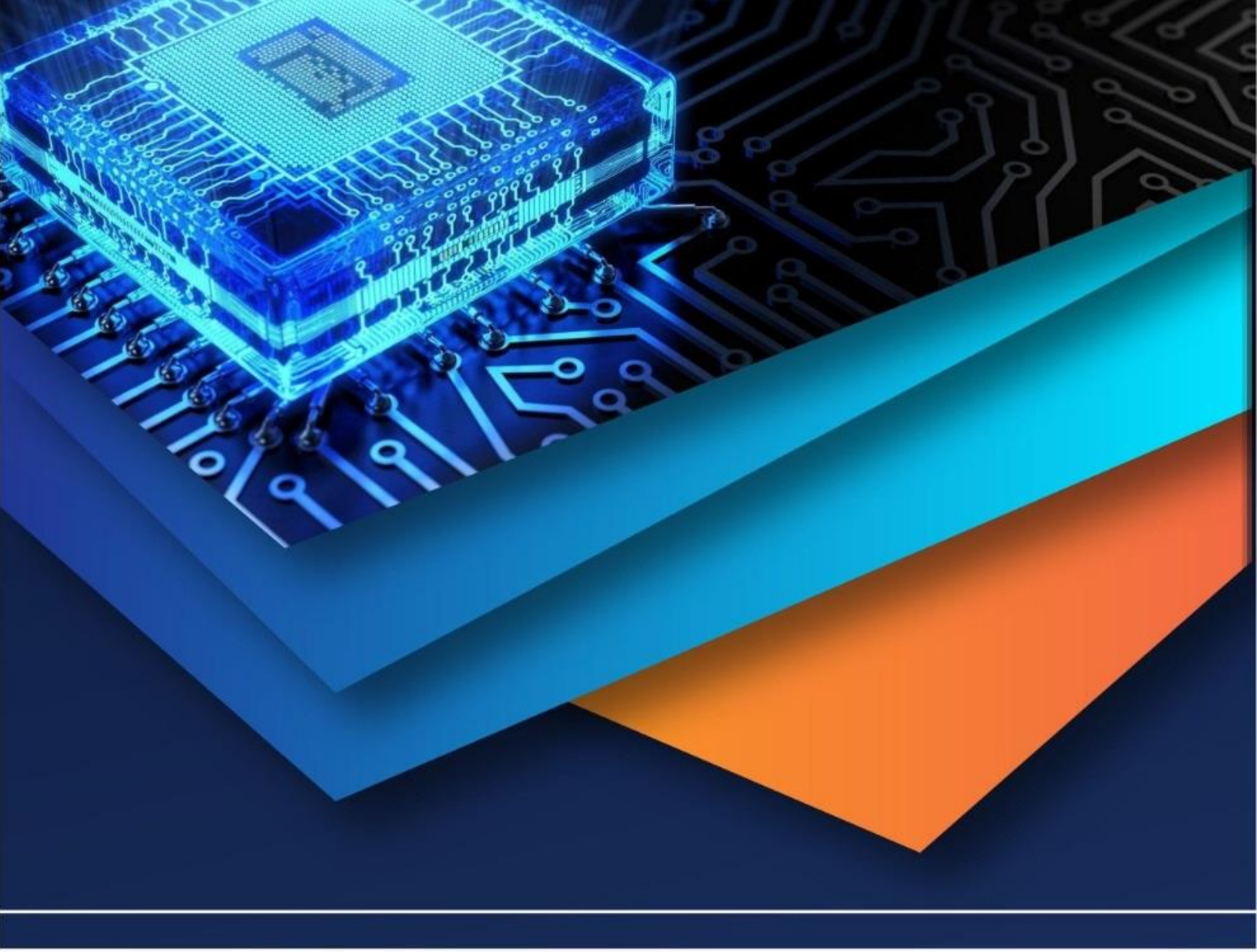
REFERENCES

[1] Liao, Q. V. (2021). Introduction to eXplainable AI (XAI) [PDF]. http://qveraliao.com/xai_tutorial.pdf
[2] 10Senses. (n.d.). Introduction to Explainable AI (Explainable Artificial Intelligence or XAI). https://10senses.com/blog/introduction-to-explainable-ai-explainable-artificial-intelligence-or-xai/
[3] TechTarget. (n.d.). What is explainable AI? | Definition from TechTarget. https://www.techtarget.com/whatis/definition/explainable-AI-XAI
[4] Heckman, C. (2022, January 17). What is Explainable AI? - SEI Blog. https://insights.sei.cmu.edu/blog/what-is-explainable-ai/
[5] Tripwire. (2024, March 21). AI Transparency: Why Explainable AI Is Essential for Modern Cybersecurity. https://www.tripwire.com/state-of-security/ai-transparency-why-explainable-ai-essential-modern-cybersecurity
[6] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
[7] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural mistakes in AI models. arXiv preprint arXiv:2109.08065.
[8] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.
[9] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 80-89). IEEE.
[10] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lulu. com.
[11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
[12] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[13] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
[14] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841-887.
[15] Molnar, C. (2022). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable
[16] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42
[17] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144)
[18] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4768-4777)
[19] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
[20] Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049.
[21] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[22] Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-15).
[23] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77-91)
[24] Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 10-19)
[25] El Arras, L., Leite, J., Kern, K., Kloft, M., & Valera, I. (2022). CERTIFY: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of AI models. arXiv preprint arXiv:2202.07845
[26] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115
[27] Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. Minds and Machines, 29(4), 495-510.
[28] Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press Karimi, A. H., Schölkopf, B., & Valera, I. (2020). Algorithmic recourse: From counterfactual explanations to interventions. arXiv preprint arXiv:2002.06278.
[29] Mahajan, D., Tan, C., & Sharma, A. (2019). Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv preprint arXiv:1912.03277.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[30] Puiutta, E., & Veith, E. M. S. P. (2020). Explainable reinforcement learning: A survey. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 77-95). Springer, Cham
[31] Huang, S. H., Held, D., Abbeel, P., & Mustafa, W. (2019). Interpretable and data-efficient reinforcement learning for robotics. arXiv preprint arXiv:1906.04950.
[32] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.
[33] Kulesza, T., Stumpfhauser, D., Burnett, M., Wong, W. K., Riche, Y., Morre, A., ... & Oberst, I. (2013). Explanatory debugging: Supporting end-user understanding of machine learning. In Explanation in Computing (pp. 21-60). Springer, Cham.
[34] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
[35] Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Distill-and-explain: Distilling model explanations for AI interpreting systems. arXiv preprint arXiv:1811.07619.
[36] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In International Conference on Machine Learning (pp. 3319-3328).
[37] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[38] Mohseni, S., Zarei, N., & Ragan, E. D. (2020). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems, 11(3-4), 1-45.
[39] European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence. Retrieved from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence
[40] IEEE. (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf
[41] Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.
[42] Puiutta, E., & Veith, E. M. S. P. (2020). Explainable reinforcement learning: A survey. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 77-95). Springer, Cham.
[43] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)