Journal of Frontiers in Multidisciplinary Research

Explainable AI in Robotics: A Critical Review and Implementation Strategies for Transparent Decision-Making

Abiodun Sunday Adebayo [1*], **Olanrewaju Oluwaseun Ajayi** [2], **Naomi Chukwurah** [3]
[1] University of Staffordshire, United Kingdom
[2] University of the Cumberlands, USA
[3] Independent Researcher, USA

* Corresponding Author: **Abiodun Sunday Adebayo**

Article Info

E-ISSN: 3050-9726
P-ISSN: 3050-9718
Volume: 05
Issue: 01
January-June 2024
Received: 21-11-2023
Accepted: 13-12-2023
Published: 08-01-2024
Page No: 26-32

Abstract
The rapid advancement of AI-driven robotic systems has introduced significant challenges related to transparency and trust, particularly in safety-critical applications. This review paper critically examines the current approaches to Explainable AI (xAI) in robotics, emphasizing the inherent trade-offs between performance and transparency. While high-performance AI models are essential for complex robotic tasks, their opacity often undermines trust and limits adoption. To address this, the paper proposes a comprehensive framework for implementing xAI in robotics, including strategies such as modular architecture, hybrid models, and human-centered design. The paper also discusses key design considerations and evaluation metrics that ensure a balance between interpretability and operational effectiveness. Finally, the paper reflects on the implications of these strategies for the future of robotics. It suggests avenues for further research to enhance the integration of xAI, aiming to create more trustworthy and reliable robotic systems.

DOI: https://doi.org/10.54660/.IJFMR.2024.5.1.26-32

Keywords: Explainable AI (xAI), Robotics, Transparency, Performance trade-offs, Safety-critical applications

1. Introduction

The integration of Artificial Intelligence (AI) into robotic systems has revolutionized various industries, ranging from healthcare and manufacturing to autonomous vehicles and military applications (Rashid, Kausik, Al Hassan Sunny, & Bappy, 2023). AI-driven robotics can perform complex tasks with high autonomy, efficiency, and precision. However, despite these advancements, one of the critical challenges hindering the widespread adoption of AI in robotics, particularly in safety-critical applications, is the lack of transparency in decision-making processes. The opaque nature of AI models, especially those based on deep learning, creates a "black box" effect, where even the developers and operators of these systems may not fully understand how decisions are made. This lack of transparency raises significant safety, reliability, and accountability concerns (Taj & Zaman, 2022).

In safety-critical domains such as healthcare, autonomous driving, and military operations, erroneous decisions can be catastrophic, leading to loss of life, severe injury, or substantial financial and reputational damage. For instance, in the case of autonomous vehicles, an AI system's inability to explain its decision-making process can result in a loss of trust among users, regulators, and the general public (Avraham & Porat, 2023). Similarly, in healthcare, a robot-assisted surgery system that cannot provide a rationale for its actions may face resistance from medical professionals and patients alike. Therefore, transparency in AI-driven robotic systems is not merely a technical requirement but a fundamental prerequisite for building trust, ensuring safety, and facilitating broader adoption in critical applications (Alzubaidi *et al*, 2023).

The primary objective of this paper is to critically review the current approaches to Explainable AI (xAI) in robotics and to propose implementation strategies that balance performance with transparency. Explainable AI refers to a set of methodologies and techniques that make the decision-making processes of AI models more understandable to humans. In robotics, xAI aims to elucidate how and why a robotic system arrives at a particular decision or action, enhancing interpretability and trustworthiness. However, the pursuit of explain ability often comes at the cost of performance, as more interpretable models tend to be less complex and, consequently, less powerful than their opaque counterparts.

This paper will explore the delicate trade-offs between performance and transparency in AI-driven robotic systems. It will examine whether achieving a balance that allows for high-performing yet interpretable models is possible or if compromises must be made in one area to enhance the other. By analyzing existing xAI techniques and their application in robotics, this paper seeks to identify gaps in current methodologies and offer practical strategies for improving both the transparency and effectiveness of AI in robotic systems. Ultimately, the paper aims to contribute to developing more interpretable and trustworthy robotic systems, which are essential for their safe deployment in critical real-world applications.

2. Current approaches to explainable ai in robotics

2.1 Overview of explainable AI

Explainable AI (xAI) represents a critical frontier in the development of artificial intelligence, especially within robotics. xAI encompasses a range of techniques and methodologies designed to make AI systems more interpretable, allowing human users to understand and trust the decisions made by these systems (Tjoa & Guan, 2020). The need for explainability becomes particularly acute in robotics, where AI-driven systems are often deployed in complex, dynamic environments. Robots are increasingly tasked with making autonomous decisions that can have significant consequences—whether in healthcare, autonomous driving, manufacturing, or military operations. As such, the ability to explain how and why a robot arrived at a specific decision is crucial for ensuring safety, accountability, and user trust (Saeed & Omlin, 2023).

The principles of xAI are grounded in the idea that AI systems should not be "black boxes." However, they should instead provide insights into their decision-making processes. The goals of xAI in robotics include enhancing transparency, improving user trust, and enabling better oversight of AI-driven decisions (Chamola *et al*, 2023). This is particularly important in applications where errors could lead to harmful outcomes, as an explainable system can provide justifications for its actions, making it easier to identify and correct potential flaws. Additionally, explainability facilitates compliance with regulatory requirements and ethical standards, which are becoming increasingly stringent as AI systems become more pervasive (Díaz-Rodríguez *et al*, 2023).

2.2 Techniques and methods

Several techniques and methods have been developed to achieve explainability in AI systems, and many of these have been adapted for use in robotics. One of the most common approaches is model interpretability, which involves designing AI models to make their internal workings understandable to humans (Hong, Hullman, & Bertini, 2020). For example, decision trees and rule-based systems are inherently interpretable because they allow users to trace the decision-making path from inputs to outputs. However, these models are often too simplistic to handle the complexities of real-world robotic tasks, leading researchers to explore more sophisticated techniques.

Feature importance is another widely used technique in xAI. This method identifies which features (or inputs) are most influential in determining the output of an AI model. In robotics, this could involve determining which sensor readings or environmental cues are most critical in guiding a robot's actions. Tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been developed to assess feature importance in complex models, including those used in robotics (Brdnik, Podgorelec, & Šumak, 2023; Dwivedi *et al*, 2023).

Saliency maps are another popular technique integral to many robotic systems, particularly in computer vision. A saliency map highlights areas of an image that are most influential in the decision-making process of a model. For instance, in a robot designed for object detection and manipulation, a saliency map can reveal which parts of the visual input the model focuses on when identifying objects. This can help operators understand and trust the robot's decisions, as they can see what it "sees" and why it makes certain choices (Gupta, Seal, Prasad, & Khanna, 2020).

Post-hoc explanations are also crucial in xAI, particularly for inherently opaque models, such as deep neural networks. Post-hoc techniques provide explanations after a decision has been made rather than embedding interpretability directly into the model. For example, one might use a simpler, interpretable model to approximate the behavior of a complex neural network, thereby providing insights into how the original model functions. In robotics, this approach can be useful for understanding decisions made by autonomous systems operating in real-time and unpredictable environments (Kenny & Keane, 2021). Additionally, explainability frameworks have been proposed to incorporate xAI techniques into robotic systems systematically. These frameworks provide a structured approach to building, testing, and deploying explainable robotic systems, ensuring transparency throughout the development lifecycle. For instance, the XRL (eXplainable Reinforcement Learning) framework integrates explainability into the reinforcement learning process. It is commonly used in robotics to enable robots to learn from their environment and improve their performance over time. By incorporating xAI techniques at each stage of learning, such frameworks help ensure that the resulting models are both effective and interpretable (Dazeley, Vamplew, & Cruz, 2023).

2.3 Challenges and limitations

Despite the progress in xAI, several challenges and limitations persist, particularly in robotics. One of the primary challenges is the trade-off between model complexity and explainability. More complex models, such as deep neural networks, tend to be more accurate and capable of handling the intricate tasks required in robotics. However, these models are often less interpretable, making understanding how they arrive at their decisions difficult. Simplifying the model to enhance interpretability can lead to a loss of performance, which is problematic in applications where precision and reliability are critical.

Another significant challenge is the dynamic and unpredictable nature of robotic environments. Robots often operate in real-world settings where conditions can change rapidly and unexpectedly. Ensuring that AI models can explain their decisions in such environments is difficult, as the explanations must account for a wide range of variables and potential scenarios. This complexity can make it challenging to generate accurate and comprehensible explanations to human users. The real-time requirements of many robotic systems also pose a limitation for xAI. Robots often must make decisions in milliseconds, leaving little time to generate detailed explanations. This necessitates the development of xAI techniques that can operate efficiently within strict time constraints without compromising the quality of the explanations provided (Kumar, Rajesh, Ramachandran, & Gupta, 2022).

Moreover, there is a lack of standardized evaluation metrics for xAI in robotics. While various techniques have been proposed, there is no consensus on how to measure the effectiveness of these techniques in improving transparency and trust. This lack of standardization makes it difficult to compare different approaches and to determine which methods are most suitable for specific robotic applications (Abhilashi, Singh, & Patil, 2023). Finally, the context-specific nature of explanations adds another layer of complexity. What constitutes a satisfactory explanation can vary significantly depending on the user's expertise, the specific application, and the level of risk involved. For example, a detailed technical explanation may be necessary for engineers but too complex for end-users or non-specialists. Developing xAI techniques that can tailor explanations to different audiences without losing accuracy is an ongoing challenge (Saeed & Omlin, 2023).

3. Trade-offs between performance and transparency

3.1 Performance vs. explainability

The trade-off between performance and explainability is a central challenge in designing and deploying AI-driven robotic systems. High-performing AI models, particularly those based on deep learning and other complex algorithms, are often characterized by their ability to process vast amounts of data and make decisions with remarkable accuracy. However, the complexity that enables these models to achieve such high performance makes them less interpretable, turning them into "black boxes" whose decision-making processes are difficult to understand, even for experts. On the other hand, models that prioritize explainability, such as simpler rule-based systems or decision trees, tend to be more transparent but may sacrifice accuracy and robustness, particularly in complex or unpredictable environments (Linardatos, Papastefanopoulos, & Kotsiantis, 2020).

In robotics, this trade-off is particularly significant due to the nature of the tasks these systems are expected to perform. For instance, in autonomous vehicles, AI systems must process real-time data from multiple sensors, make split-second decisions, and navigate through dynamic environments with high levels of uncertainty. Deep learning models are well-suited to this task because of their ability to learn from large datasets and make precise decisions in complex situations. However, the lack of transparency in these models poses a risk. Suppose the system makes an unexpected or incorrect decision. In that case, it can be challenging to understand why this happened, making it difficult to prevent similar errors in the future (Tofangchi, Hanelt, Marz, & Kolbe, 2021; Williams & Yampolskiy, 2021).

Conversely, suppose a robotic system is designed with simpler, more interpretable models. In that case, its decisions can be easily traced and understood. However, these models might not be able to handle the same level of complexity or make decisions with the same speed and accuracy as more sophisticated models. This can lead to suboptimal performance, particularly in scenarios where the robot must operate autonomously and handle a wide range of tasks without human intervention. Thus, designers and engineers must carefully consider the specific requirements of the application when deciding how to balance performance and explainability in their systems (Arnold, Kasenberg, & Scheutz, 2021).

3.2 Impact on safety and trust

The trade-offs between performance and explainability have profound implications for safety and trust in robotic systems, particularly in safety-critical applications where the consequences of failure can be severe. In such contexts, understanding and trusting the decisions made by AI-driven robots is paramount. For example, robotic systems are increasingly used in surgeries, diagnostics, and patient care in healthcare. Suppose a robot makes a decision that affects a patient's treatment. In that case, medical professionals must understand the reasoning behind that decision. A lack of transparency can lead to mistrust and reluctance to adopt robotic systems, even if they offer superior performance in terms of speed and accuracy (Vorm & Combs, 2022).

Similarly, the trade-off between performance and explainability in autonomous vehicles directly impacts public safety and trust. Autonomous vehicles rely on AI to interpret sensor data, predict the behavior of other road users, and make real-time driving decisions. While high-performance AI models can significantly reduce the likelihood of accidents, the opacity of these models means that when accidents occur, it can be difficult to determine the cause. This not only complicates the process of assigning liability but also undermines public trust in autonomous driving technology. Users and regulators are more likely to accept and trust autonomous vehicles if they understand how they make decisions, particularly in critical situations.

Moreover, the lack of explainability in high-performance models can hinder the development of safety regulations and standards. Regulators must understand how AI-driven robotic systems function to establish guidelines ensuring their safe deployment. If these systems' decision-making processes are opaque, assessing their safety and efficacy becomes challenging. This can lead to delays in the approval and adoption of new technologies, limiting the potential benefits of AI in robotics. The impact on trust is not limited to end-users and regulators but also extends to developers and operators of robotic systems. Engineers need to be able to diagnose and fix issues when they arise, and this is far more difficult when the AI models in use are not interpretable. Identifying and rectifying errors becomes time-consuming and complex without explaining a robot's actions. This can lead to operational inefficiencies and increase the risk of errors, eroding trust in the technology (Fisher *et al*, 2021).

3.3 Examples and analysis

Various approaches have been developed to address the trade-offs between performance and transparency, each with strengths and weaknesses. One common strategy is to use hybrid models that combine interpretable components with high-performance machine learning techniques. For instance, a robot might use a rule-based system to handle routine tasks where transparency is crucial while relying on a deep learning model for more complex tasks that require higher performance. This approach allows developers to balance the need for explainability in certain contexts with the need for high performance in others. However, integrating different types of models can be challenging, and ensuring smooth transitions between them can add complexity to the system (Masís, 2021).

Another approach involves using post-hoc explanation techniques, which explain the decisions made by complex models after the fact. Techniques such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are commonly used to generate human-readable explanations for the outputs of deep learning models. These methods allow developers and users to understand how the model arrived at a particular decision, even if it is not inherently interpretable. While post-hoc explanations can enhance transparency, they do not fundamentally alter the opacity of the underlying model, and there is always the risk that the explanations provided may not fully capture the model's decision-making process (Dwivedi *et al*, 2023; Zafar & Khan, 2021).

Sometimes, efforts are made to simplify complex models without significantly compromising performance. For example, techniques such as model pruning, which removes unnecessary parameters from a deep learning model, can make the model more interpretable while maintaining high accuracy. Similarly, distillation methods involve training a simpler model to mimic the behavior of a more complex model. The simpler model, which is easier to interpret, can then be used to generate explanations for the decisions made by the more complex model. These approaches can be effective, but they often involve a trade-off between the degree of simplification and the retention of the model's original performance (Yeom *et al*, 2021).

Explainability frameworks also play a crucial role in balancing performance and transparency. For instance, using reinforcement learning in robotics often involves complex decision-making processes that are difficult to explain. However, developers can create models that offer a better balance between performance and transparency by incorporating explainability into the reinforcement learning framework itself—such as through interpretable reward structures or the integration of human-in-the-loop systems. These frameworks help ensure that robots can learn from their environment in a way that is both effective and understandable to humans (Sado, Loo, Liew, Kerzel, & Wermter, 2023; Zelvelder, Westberg, & Främling, 2021).

4. Implementation strategies for balancing performance and transparency

4.1 Framework for implementation

Implementing Explainable AI (xAI) in robotics necessitates a well-structured framework that balances the need for high performance with the demand for transparency. The complexity of robotic systems, combined with the often-critical nature of their applications, makes it essential to design AI models that are both effective and interpretable. Therefore, a robust implementation framework should include several key components: an iterative development process, modular architecture, and human-centered design (Liao & Varshney, 2021).

The iterative development process involves continuously refining AI models through development, testing, and feedback cycles. This approach ensures that performance and transparency are incrementally improved with each iteration. The process begins with developing a baseline model that prioritizes either performance or transparency, depending on the specific application. As the model is tested in real-world conditions, feedback is gathered from various stakeholders, including developers, end-users, and regulators. This feedback is then used to adjust the model, improving transparency without significantly compromising performance. Over time, this iterative approach helps balance these two often competing objectives (Haakman, Cruz, Huijgens, & Van Deursen, 2021).

A modular architecture is also critical to this framework. By designing robotic systems with modular AI components, developers can isolate different functionalities—perception, decision-making, and action execution—and apply xAI techniques where they are most needed. For instance, a robot's perception module might use deep learning models for high performance. In contrast, its decision-making module could incorporate rule-based systems or other interpretable models to enhance transparency. This modular approach allows for flexibility in balancing performance and transparency across different parts of the system, ensuring that critical decisions are understandable without compromising the overall effectiveness of the robot (Sado *et al*, 2023; Vice & Khan, 2022).

Human-centered design is the third key component of the implementation framework. This approach emphasizes the importance of considering the needs and expectations of end-users throughout the design process. By involving users in the development of xAI systems, developers can ensure that the explanations provided by the AI are meaningful and relevant to those who interact with the robot. This might involve tailoring explanations to the user's level of expertise and providing more detailed information to technical users while offering simpler, more intuitive explanations to non-experts. Human-centered design improves the usability of robotic systems and fosters trust by ensuring that users feel confident in the robot's decisions (Göttgens & Oertelt-Prigione, 2021; Sanneman & Shah, 2022).

4.2 Design Considerations

When integrating xAI techniques into robotic systems, several design considerations and best practices must be considered to maintain high performance. One of the most important considerations is the selection of appropriate xAI techniques based on the specific application and the nature of the tasks the robot will perform. For instance, in applications where real-time decision-making is critical, such as autonomous driving or emergency response, it may be necessary to prioritize performance while using lightweight xAI techniques that provide quick, albeit less detailed, explanations. On the other hand, in contexts where understanding the decision-making process is more important than speed—such as in medical diagnostics—more comprehensive xAI techniques, such as model interpretability methods or post-hoc explanations, might be preferred (Islam, Ahmed, Barua, & Begum, 2022).

Another key design consideration is the complexity of the AI model. While more complex models offer better performance, they are also more challenging to explain. Developers should, therefore, aim to simplify models where possible without sacrificing the necessary level of performance. Techniques such as pruning, where unnecessary parameters are removed from a model, or distillation, where a simpler model is trained to replicate the behavior of a more complex one, can help achieve this balance. Additionally, combining different models—such as using an interpretable model for high-stakes decisions and a more complex model for routine tasks—can allow for high performance and transparency within the same system (Hoefler, Alistarh, Ben-Nun, Dryden, & Peste, 2021).

The integration of user feedback is another critical consideration. Throughout the design process, developers should actively seek input from users to understand their needs and preferences regarding transparency. This feedback can guide the selection and refinement of xAI techniques, ensuring that the explanations provided by the system are aligned with user expectations. Moreover, incorporating user feedback into the iterative development process can help identify areas where the model's transparency might be improved without detracting from its performance (Springer & Whittaker, 2020).

Finally, developers should consider the context in which the robotic system will be deployed. Different environments and use cases may require different approaches to balancing performance and transparency. For example, a robot operating in a controlled industrial setting might prioritize performance with minimal need for transparency, as its actions are closely monitored and predictable. Conversely, a robot in a public space, interacting with a diverse range of people, may need to prioritize transparency to gain trust and ensure safe interactions. By carefully considering the deployment context, developers can make informed decisions about integrating xAI techniques into their systems (Asatiani *et al*, 2021).

4.3 Evaluation Metrics

To effectively balance performance and transparency in xAI implementations, it is essential to establish clear metrics for evaluating the success of these efforts. These metrics should assess both the interpretability of the AI model and its operational performance, providing a comprehensive view of the system's overall effectiveness. One of the primary metrics for evaluating interpretability is explanation fidelity. This metric assesses how accurately the explanations of the xAI techniques reflect the AI model's underlying decision-making process. High explanation fidelity ensures that the explanations are understandable and truthful representations of the model's operations. This metric can be measured by comparing the outcomes of the original model with those of a simplified, interpretable model designed to mimic its behavior. The closer the two models' outputs, the higher the fidelity of the explanation (Stevens & De Smedt, 2024).

User satisfaction is another crucial metric for assessing interpretability. This involves gathering feedback from end-users regarding the clarity and usefulness of the explanations provided by the robot. Surveys, interviews, and usability testing can all be used to gauge how well users understand the robot's decisions and whether they trust the system as a result. High user satisfaction indicates that the xAI techniques have been successfully implemented and the system is transparent and trustworthy (Brdnik *et al*, 2023).

In terms of operational performance, accuracy and efficiency remain the primary metrics. Accuracy measures how well the AI model performs its intended tasks. At the same time, efficiency assesses the speed and resource usage of the model. It is essential to monitor these metrics to ensure that efforts to improve transparency do not significantly degrade the robot's performance. In applications where real-time decision-making is critical, efficiency may take on added importance, with developers needing to balance providing explanations and maintaining quick response times (Baccour *et al*, 2022).

Another important performance-related metric is robustness, which evaluates the AI model's ability to perform reliably under varying conditions. A robust model should maintain accuracy and efficiency even when faced with unexpected inputs or environmental changes. Ensuring that xAI techniques do not compromise the robustness of the model is crucial, especially in safety-critical applications where consistent performance is vital (Chander, John, Warrier, & Gopalakrishnan, 2024). Finally, compliance with regulatory standards can serve as an indirect metric for evaluating performance and transparency. Many industries, particularly those involving safety-critical applications like healthcare or autonomous driving, have specific regulations that dictate the level of transparency required for AI systems. Ensuring that the robotic system meets these regulatory requirements enhances trust and safety and demonstrates that the system has successfully balanced performance with transparency (Liu, 2023; Topcu *et al*, 2020).

5. Conclusion and future directions

5.1 Conclusion

This paper has critically examined the intersection of Explainable AI (xAI) and robotics, highlighting the ongoing challenges and trade-offs between performance and transparency in AI-driven robotic systems. The review of current xAI approaches revealed that while significant progress has been made in developing techniques that enhance the interpretability of AI models, achieving a balance between high performance and transparency remains a complex task. Strategies such as modular architecture, hybrid models, and post-hoc explanation techniques were identified as effective methods for integrating xAI into robotic systems. Furthermore, the importance of an iterative development process, human-centered design, and robust evaluation metrics was emphasized to ensure that AI-driven robots are effective and understandable.

The findings of this paper have significant implications for the field of robotics, especially in safety-critical applications such as healthcare, autonomous driving, and industrial automation. The balance between performance and transparency is not just a technical challenge but also a critical factor in end-users and regulators' adoption and trust of robotic systems. As robots increasingly perform complex and autonomous tasks, the need for transparency becomes more pressing. Users must be able to trust these systems, particularly when their safety is at stake. By implementing xAI techniques, developers can create robotic systems that are capable of high performance and provide clear and understandable explanations for their actions, thus enhancing user trust and facilitating wider adoption in critical applications.

Moreover, the emphasis on human-centered design within the proposed implementation strategies reflects a shift towards more user-friendly robotic systems. By tailoring explanations to the needs and expertise of different users, robotic systems can become more accessible and less intimidating, encouraging broader acceptance across various sectors. Additionally, the integration of xAI will likely influence the development of industry standards and regulations, which increasingly require AI systems to be transparent and accountable. Therefore, the strategies proposed in this paper address the current technical challenges and align with the broader trend toward responsible and ethical AI in robotics.

5.2 Future research directions

While the proposed strategies offer a roadmap for integrating xAI into robotic systems, there are still several areas where further research is needed to address existing limitations and explore new approaches. One important area for future research is the development of more sophisticated hybrid models that seamlessly integrate high-performance and interpretable components. These models could leverage advancements in machine learning, such as reinforcement learning combined with interpretable decision trees, to create powerful and transparent systems.

Another promising avenue for research is the exploration of real-time xAI techniques that can provide explanations as the robot operates rather than after decisions have been made. This would be particularly useful in dynamic environments where real-time decision-making is crucial, such as autonomous vehicles or emergency response robots. Developing such techniques would require advancements in computational efficiency to ensure that explanations do not slow down the robot's operations.

Finally, research should also focus on the user experience of xAI in robotics. Understanding how users perceive and interact with AI-driven robots and what types of explanations are most effective in building trust will be crucial for refining xAI techniques. This could involve interdisciplinary studies that combine insights from psychology, human-computer interaction, and AI, aiming to develop more intuitive and user-friendly interfaces for robotic systems.

6. References

1. Abhilashi L, Singh V, Patil SK. New Transportation Engineering Technology. GEH Press; 2023.
2. Alzubaidi L, Al-Sabaawi A, Bai J, Dukhan A, Alkenani AH, Al-Asadi A, *et al* Towards risk-free trustworthy artificial intelligence: Significance and requirements. International Journal of Intelligent Systems. 2023;2023(1):4459198.
3. Arnold T, Kasenberg D, Scheutz M. Explaining in time: Meeting interactive standards of explanation for robotic systems. ACM Transactions on Human-Robot Interaction (THRI). 2021;10(3):1–23.
4. Asatiani A, Malo P, Nagbøl PR, Penttinen E, Rinta-Kahila T, Salovaara A. Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. Journal of the Association for Information Systems. 2021;22(2):325–52.
5. Avraham R, Porat A. The dark side of insurance. Review of Law & Economics. 2023;19(1):13–45.
6. Baccour E, Mhaisen N, Abdellatif AA, Erbad A, Mohamed A, Hamdi M, *et al* Pervasive AI for IoT applications: A survey on resource-efficient distributed artificial intelligence. IEEE Communications Surveys & Tutorials. 2022;24(4):2366–418.
7. Brdnik S, Podgorelec V, Šumak B. Assessing perceived trust and satisfaction with multiple explanation techniques in XAI-enhanced learning analytics. Electronics. 2023;12(12):2594.
8. Chamola V, Hassija V, Sulthana AR, Ghosh D, Dhingra D, Sikdar B. A review of trustworthy and explainable artificial intelligence (XAI). IEEE Access. 2023;11:1–25.
9. Chander B, John C, Warrier L, Gopalakrishnan K. Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. ACM Computing Surveys. 2024;57(3):1–30.
10. Dazeley R, Vamplew P, Cruz F. Explainable reinforcement learning for broad-XAI: A conceptual framework and survey. Neural Computing and Applications. 2023;35(23):16893–916.
11. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, de Prado ML, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion. 2023;99:101896.
12. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, *et al* Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys. 2023;55(9):1–33.
13. Fisher M, Cardoso RC, Collins EC, Dadswell C, Dennis LA, Dixon C, *et al* An overview of verification and validation challenges for inspection robots. Robotics. 2021;10(2):67.
14. Göttgens I, Oertelt-Prigione S. The application of human-centered design approaches in health research and innovation: A narrative review of current practices. JMIR mHealth and uHealth. 2021;9(12):e28102.
15. Gupta AK, Seal A, Prasad M, Khanna P. Salient object detection techniques in computer vision—A survey. Entropy. 2020;22(10):1174.
16. Haakman M, Cruz L, Huijgens H, Van Deursen A. AI lifecycle models need to be revised: An exploratory study in fintech. Empirical Software Engineering. 2021;26(5):95.
17. Hoefler T, Alistarh D, Ben-Nun T, Dryden N, Peste A. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. Journal of Machine Learning Research. 2021;22(241):1–124.
18. Hong SR, Hullman J, Bertini E. Human factors in model interpretability: Industry practices, challenges, and needs. Proceedings of the ACM on Human-Computer Interaction. 2020;4(CSCW1):1–26.
19. Islam MR, Ahmed MU, Barua S, Begum S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences. 2022;12(3):1353.
20. Kenny EM, Keane MT. Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. Knowledge-Based Systems. 2021;233:107530.
21. Kumar A, Rajesh T, Ramachandran M, Gupta D. Role of explainable edge AI to resolve real-time problems. In: Explainable Edge AI: A Futuristic Computing Perspective. Springer; 2022. p. 101–16.
22. Liao QV, Varshney KR. Human-centered explainable AI (XAI): From algorithms to user experiences. arXiv preprint arXiv:2110.10790; 2021.
23. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy. 2020;23(1):18.
24. Liu Y. Explainable AI techniques for transparency in autonomous vehicle decision-making. Journal of AI in

Healthcare and Medicine. 2023;3(2):114–34.

25. Masís S. Interpretable Machine Learning with Python: Learn to Build Interpretable High-Performance Models with Hands-On Real-World Examples. Packt Publishing Ltd; 2021.
26. Rashid AB, Kausik AK, Al Hassan Sunny A, Bappy MH. Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. International Journal of Intelligent Systems. 2023;2023(1):8676366.
27. Sado F, Loo CK, Liew WS, Kerzel M, Wermter S. Explainable goal-driven agents and robots: A comprehensive review. ACM Computing Surveys. 2023;55(10):1–41.
28. Saeed W, Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems. 2023;263:110273.
29. Sanneman L, Shah JA. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. International Journal of Human–Computer Interaction. 2022;38(18-20):1772–88.
30. Springer A, Whittaker S. Progressive disclosure: When, why, and how do users want algorithmic transparency information? ACM Transactions on Interactive Intelligent Systems (TiiS). 2020;10(4):1–32.
31. Stevens A, De Smedt J. Explainability in process outcome prediction: Guidelines to obtain interpretable and faithful models. European Journal of Operational Research. 2024;317(2):317–29.
32. Taj I, Zaman N. Towards industrial revolution 5.0 and explainable artificial intelligence: Challenges and opportunities. International Journal of Computing and Digital Systems. 2022;12(1):295–320.
33. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems. 2020;32(11):4793–813.
34. Tofangchi S, Hanelt A, Marz D, Kolbe LM. Handling the efficiency–personalization trade-off in service robotics: A machine-learning approach. Journal of Management Information Systems. 2021;38(1):246–76.
35. Topcu U, Bliss N, Cooke N, Cummings M, Llorens A, Shrobe H, Zuck L. Assured autonomy: Path toward living with autonomous systems we can trust. arXiv preprint arXiv:2010.14443; 2020.
36. Vice J, Khan MM. Toward accountable and explainable artificial intelligence part two: The framework implementation. IEEE Access. 2022;10:36091–105.
37. Vorm ES, Combs DJ. Integrating transparency, trust, and acceptance: The intelligent systems technology acceptance model (ISTAM). International Journal of Human–Computer Interaction. 2022;38(18-20):1828–45.
38. Williams R, Yampolskiy R. Understanding and avoiding AI failures: A practical guide. Philosophies. 2021;6(3):53.
39. Yeom S-K, Seegerer P, Lapuschkin S, Binder A, Wiedemann S, Müller K-R, Samek W. Pruning by explaining: A novel criterion for deep neural network pruning. Pattern Recognition. 2021;115:107899.
40. Zafar MR, Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability. Machine Learning and Knowledge Extraction. 2021;3(3):525–41.
41. Zelvelder AE, Westberg M, Främling K. Assessing explainability in reinforcement learning. Paper presented at: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems; 2021.