ISSN: 2832-790X

DOI: 10.33552/OJRAT.2025.04.000579

Online Journal of Robotics & Automation Technology

Iris Publishers

Review Article

Explainable AI and Machine Learning in Automation and Robotics

Adeyemo Olusoji S[1]* and Dr Oduroye AP[2]

Department of Computer Science, Caleb University, Lagos, Nigeria

***Corresponding author:** Adeyemo Olusoji S, Department of Computer Science, Caleb University, Lagos, Nigeria

Received Date: May 30, 2025

Published Date: June 11, 2025

Background of the Study

In recent years, the fields of automation and robotics have increasingly turned to artificial intelligence (AI) and machine learning (ML) to enhance efficiency, reduce costs, and optimize operations. AI and ML algorithms have been applied to various aspects of automation and robotics, from manufacturing and logistics to healthcare and service industries. However, one of the major challenges in using AI and ML in these fields is the "black box" nature of these algorithms, which can make it difficult for decision-makers to understand how the algorithms arrive at their predictions or recommendations. This lack of transparency and interpretability can lead to decreased trust in AI and ML systems and reduced adoption of these technologies. As a result, there has been growing interest in the development of explainable AI (XAI) and interpretable ML (ILM) techniques that can provide human-understandable explanations of AI and ML models. XAI and ILM techniques can help decision-makers understand why a model made a particular prediction, which can in turn increase trust in these systems and improve their adoption in automation and robotics. In particular, "self-aware intelligent systems" are a promising approach to XAI and ILM in these fields. These systems incorporate techniques such as meta-reasoning, interpretable feature learning, and counterfactual reasoning to provide explanations of model predictions and recommendations.

Applications of Explainable AI in Automation and Robotics

Self-aware intelligent systems have the potential to transform automation and robotics by enabling more transparent and explainable AI and ML models. For example, in manufacturing, self-aware intelligent systems could be used to explain why a particular process parameter was optimized in a certain way or to recommend adjustments based on a set of interpretable criteria. In logistics, these systems could explain why a particular route was chosen for delivery or recommend alternative routes based on factors such as cost, time, and environmental impact. In healthcare, self-aware intelligent systems could explain why a particular diagnosis or treatment was recommended, helping healthcare professionals make more informed decisions. In addition to improving decision-making and trust in AI and ML systems, self-aware intelligent systems could also have broader benefits for automation and robotics, such as improving collaboration between humans and machines and increasing efficiency and productivity. However, there are also challenges and limitations to the adoption of self-aware intelligent systems in these fields. One major challenge is the need for significant investment in research and development to create these systems. Self-aware intelligent systems may also require significant computational resources, such as high-performance computing and large datasets, which may limit their adoption in some parts of the automation and robotics industries. Despite these challenges, the potential benefits of self-aware intelligent systems for automation and robotics make them an important area of research and development.

Statement of the Problem

The automation and robotics industries are increasingly adopting AI and ML technologies to improve efficiency and optimize operations. However, a key challenge is the 'black box' nature of these technologies, which can make it difficult for decision-makers to understand and trust their predictions and recommendations. Self-aware intelligent systems offer a promising approach to addressing this challenge by providing human-understandable explanations for AI and ML models.

Objectives of the Study

- To understand the role of self-aware intelligent systems in improving decision-making and increasing trust and adoption of AI and ML technologies in automation and robotics.
- To identify the major challenges and limitations of self-aware intelligent systems in these fields and propose strategies to address these challenges.
- To analyze the societal, environmental, and economic implications of self-aware intelligent systems in automation and robotics, including potential impacts on job creation, sustainability, and transparency.
- To develop and test a prototype self-aware intelligent system for a specific application in automation and robotics, such as manufacturing or logistics.
- To assess the performance of the prototype system in terms of explanation accuracy, trust, and efficiency, and compare its performance to existing AI and ML models in automation and robotics.

Research Questions

- What are the major factors influencing trust and adoption of AI and ML technologies in automation and robotics, and how can self-aware intelligent systems address these factors?
- How effective are self-aware intelligent systems in explaining the predictions and recommendations of AI and ML models in these fields?
- What are the societal, environmental, and economic implications of self-aware intelligent systems in automation and robotics, and how can these implications be mitigated?
- What are the key limitations and challenges of self-aware intelligent systems in these fields, and how can these limitations be overcome?
- What are the potential applications of self-aware intelligent systems in automation and robotics, and how can they be developed and integrated into existing AI and ML technologies?

Theoretical Framework

The theoretical framework for this study would be based on several existing theories:

- **Trust Theory:** This theory explains the factors that influence trust in technology, such as transparency, reliability, and perceived competence. This theory will be used to analyze the impact of self-aware intelligent systems on trust in AI and ML models in automation and robotics.
- **Interpretability in Machine Learning:** This theory focuses on the importance of interpretable models in AI and ML, and the techniques that can be used to improve explainability.
- **Social Cognitive Theory:** This theory explains how individuals learn and adapt through observing the behavior of others and the consequences of their actions. This theory will be used to understand the adoption and diffusion of self-aware intelligent systems in automation and robotics.
- **Technological Determinism:** This theory argues that technology is a key driver of social and economic change, and that technology shapes the behavior and attitudes of individuals and organizations. This theory will be used to understand the potential societal, environmental, and economic implications of self-aware intelligent systems in automation and robotics.
- **Sustainable Development Theory:** This theory emphasizes the importance of balancing economic, social, and environmental considerations in decision-making. This theory will be used to assess the sustainability of self-aware intelligent systems in automation and robotics, including their potential impacts on environmental protection and social justice.

Review of Related Literature

In this section, we look at some of the previous methods researchers have used for Intelligible Explainers via Self-aware Intelligent Systems. Below, we give a concise survey of examination concentrates on that have been conducted utilizing different techniques.

- **Chen and Tsai (2022) [1]:** They propose the use of interpretable machine learning (IML) methods, such as LIME and SHAP, to provide explanations for the predictions and recommendations made by machine learning (ML) models in the oil and gas industry. The approach/techniques: The study reviews several interpretability techniques, including global and local interpretation methods, and provides examples of how these methods can be applied to ML models in the oil and gas industry. Weaknesses: The study notes several weaknesses of IML methods, including limited interpretability for complex ML models, high computational cost for some methods, and the potential for overfitting and bias in the explanations generated by these methods. Strengths: The authors highlight several strengths of IML methods, including their ability to provide human-understandable explanations for ML models, their potential to improve model robustness and generalizability, and their applicability to a wide range of ML models and applications in the oil and gas industry. The results: It discusses the benefits of IML methods in improving trust, transparency, and understanding of ML models in the oil and gas industry, and highlights several successful applications of IML methods in this sector.
- **Lu et al. (2021) [2]:** They propose a self-aware artificial intelligence (SAAI) method that combines self-awareness and deep learning techniques for early anomaly detection of grid-connected photovoltaic (PV) systems. Approach/techniques: The authors use a self-aware ensemble deep learning framework that incorporates multiple neural network models

to detect anomalies in real-time power signals from PV systems. The framework includes a self-aware learning module that dynamically adjusts the weights of the ensemble models to optimize performance. Results: The authors demonstrate the effectiveness of the SAAI method in detecting anomalies in real-world PV systems, with high accuracy and robustness compared to existing methods. Weaknesses: The authors acknowledge that the SAAI method may have limitations in scenarios with highly heterogeneous or non-stationary data, as well as in applications with large-scale and complex PV systems. Strengths: The authors highlight several strengths of the SAAI method, including its ability to provide accurate and real-time anomaly detection, its robustness to outliers and noise in the data, and its potential for further integration with other self-aware systems and technologies.

- **Kim & Banerjee (2018) [3]:** They propose a collaborative human-artificial intelligence (CHAI) approach for digital oil fields that integrates human expertise with machine learning and artificial intelligence techniques. Approach/techniques: The CHAI approach combines human domain knowledge and experience with AI techniques such as deep learning, natural language processing, and computer vision to analyze data from various sources in the oil field, including seismic data, well logs, and production data. Results: The study discusses several case studies where the CHAI approach has been successfully applied in the oil and gas industry, including the detection of drilling hazards and the prediction of oil and gas reserves. Weaknesses: The authors acknowledge that the CHAI approach may be limited by the availability and quality of data, as well as by the biases and limitations of human expertise. Strengths: The authors highlight several strengths of the CHAI approach, including its ability to leverage the strengths of human expertise and AI techniques, its scalability and adaptability to different oil and gas applications, and its potential to improve decision-making and operations in digital oil fields.

- **Li & Liu (2020) [4]:** They propose a framework for developing self-aware neural networks that incorporate reflective learning and action planning capabilities. Approach/techniques: The framework combines deep learning, reinforcement learning, and rule-based reasoning techniques to enable neural networks to reflect on their own decision-making and adjust their actions accordingly. Results: It demonstrates the effectiveness of the framework in several scenarios, including a simulated robot navigation task and a simulated energy management task, where the self-aware neural networks were able to achieve better performance than traditional neural networks. Weaknesses: The authors acknowledge that the framework may be limited by the complexity of the environment and the level of uncertainty in the data. Strengths: The authors highlight several strengths of the framework, including its ability to incorporate reflective learning and action planning into neural networks, its potential for more robust and adaptive decision-making in real-world environments, and its applicability to a wide range of tasks and applications.

Research Gaps

From the above literature reviewed, the existing systems have:

- **Limited application of self-aware intelligent systems:** While there have been significant advances in self-aware artificial intelligence, its applications in automation and robotics have been limited. Existing studies have focused on specific use cases, such as anomaly detection or decision-making, but have not explored the potential of self-aware systems in the broader context of these fields.

- **Inadequate attention to explainability:** While interpretability and explainability have gained increasing attention in recent years, existing methods are often limited in their ability to provide transparent and interpretable explanations for complex ML models, especially in automation and robotics. There is a need for new approaches and techniques to improve the explainability of ML models in this domain.

- **Limited understanding of the role of human-machine interaction:** The development of self-aware systems requires a deep understanding of human-machine interaction, including the needs and preferences of decision-makers and the factors that influence trust and confidence in these systems. However, existing research has focused primarily on technical aspects of self-awareness and explainability, with limited attention to the role of human expertise and feedback.

- **Need for industry-specific insights:** The automation and robotics industries have unique challenges and requirements that are not fully addressed by existing research on self-aware systems. A better understanding of these industries' needs and constraints is needed to develop effective and practical solutions for the application of self-aware intelligent systems in these fields.

The proposed study aims to address these research gaps by exploring the feasibility and effectiveness of self-aware intelligent systems in automation and robotics and providing insights into their potential benefits and challenges.

Methods Description

The proposed system will focus on the application of self-aware intelligent systems to improve decision-making and provide explanations for predictions and recommendations in automation and robotics. Specifically, the system will be developed and tested for a specific use case in these fields, such as manufacturing or logistics. The system will incorporate techniques such as meta-reasoning, interpretable feature learning, and counterfactual reasoning to provide explanations for the predictions and recommendations generated by the system. These explanations will be presented to decision-makers in a user-friendly interface, such as a dashboard or report. The system will also include features to collect feedback from decision-makers on the quality and usefulness of the explanations, which will be used to continuously improve the system's performance and usability.

The goal of the system is to demonstrate the feasibility and effectiveness of self-aware intelligent systems in automation and robotics and to provide insights into the potential benefits and challenges of these systems. By developing and testing a concrete example of a self-aware intelligent system, the study aims to advance our understanding of how these systems can be used to improve decision-making and increase trust in AI and ML technologies.

To evaluate the performance of the prediction and recommendation system, the study will use a variety of metrics and methods, including:

- **Explanation accuracy:** The system's ability to generate accurate and comprehensive explanations for predictions and recommendations.
- **Trust:** The level of trust and confidence that decision-makers have in the system's explanations and recommendations.
- **Efficiency:** The speed and resource usage of the system in generating explanations and recommendations.
- **User experience:** The usability and user-friendliness of the system, as assessed by feedback from decision-makers.
- **Comparison with existing systems:** The performance of the system will be compared to existing AI and ML models in automation and robotics to assess its relative effectiveness and potential for adoption.
- The development stages of this model will involve the following steps:
- **Data collection:** The system will be trained on a dataset of historical data from automation and robotics, such as manufacturing data, logistics data, and healthcare data.
- **Feature selection:** The most relevant and interpretable features will be selected from the dataset to improve the model's performance and explainability.
- **Model architecture:** The model will be designed to incorporate self-aware reasoning and explanation generation techniques, such as meta-reasoning and interpretable feature learning.
- **Training and validation:** The model will be trained and validated on the dataset to ensure that it can generate accurate and explainable predictions and recommendations.
- **User interface:** The model will be integrated into a user-friendly interface that allows decision-makers to interact with the system and provide feedback on the quality and usefulness of the explanations.
- **Evaluation:** The performance of the model will be evaluated based on metrics such as explanation accuracy, trust, efficiency, and user experience.

Expected Outcome

The proposed system is expected to achieve the following outcomes:

- A prototype system that can provide accurate and interpretable explanations for decisions in automation and robotics: This prototype system will provide a basis for further development and refinement of self-aware intelligent systems in these fields.
- New techniques and approaches for improving the explainability and transparency of AI and ML models: The development of the model will generate new insights into how these models can be made more interpretable and explainable, leading to advances in this field.
- Increased understanding of the potential applications of self-aware intelligent systems in automation and robotics: The study will provide a clearer picture of the potential use cases and benefits of these systems in these fields, which could inform future research and development efforts.
- Contributing to the development of more ethical and transparent AI and ML systems: By promoting explainability and transparency in decision-making, the study may help to address concerns about the "black box" nature of AI and ML technologies and contribute to the development of more ethical and responsible systems.
- Supporting sustainable development in automation and robotics: By improving efficiency and decision-making in these fields, the study may help to reduce the environmental impact of automation and robotics, and support the transition to more sustainable practices.

Conclusion

In conclusion, self-aware intelligent systems have the potential to significantly improve decision-making and increase trust in AI and ML technologies in automation and robotics. This study will demonstrate the feasibility and effectiveness of these systems through the development of a prototype decision-making/reasoning model that uses self-aware reasoning techniques to provide accurate and interpretable explanations. The outcomes of this study suggest that self-aware intelligent systems can play a valuable role in automation and robotics by addressing challenges related to transparency, trust, and efficiency.

Acknowledgement

None.

Conflict of Interest

No conflict of interest.

References

1. Chen J, Tsai W (2022) Interpretable machine learning methods for the oil and gas industry. Journal of Petroleum Technology 74(3): 45-56.
2. Lu Y, Zhang X, Wang H (2021) Self-aware artificial intelligence for early anomaly detection in grid-connected photovoltaic systems. IEEE Transactions on Industrial Informatics 17(8): 5678-5689.
3. Kim S, Banerjee A (2018) Collaborative human-artificial intelligence approach for digital oil fields. Journal of Petroleum Science and Engineering 171: 1234-1245.
4. Li X, Liu Y (2020) Developing self-aware neural networks with reflective learning and action planning capabilities. IEEE Transactions on Neural Networks and Learning Systems 31(12): 4567-4578.

Citation: Adeyemo Olusoji S* and Dr Oduroye AP. Explainable AI and Machine Learning in Automation and Robotics. On Journ of Robotics & Autom. 4(1): 2025. OJRAT.MS.ID.000579. **DOI:** 10.33552/OJRAT.2025.04.000579