Explainable AI for Transparent Safety Decisions in Intelligent Safety Instrumented Systems

Waqhas Najhi Syed[1]*, Shubhashri Nikhil Waghmare[2], Love Kumar Thawait[3], Ansari Darakshan Javed[4] and Venkata Rama Rao Bommina[5]

[1]Project Manager, Syed Musab Industrial Services Company, Jeddah, Saudi Arabia
*Corresponding Email Id: syednajhi@gmail.com
[2]Associate Professor, Institute of Management and Entrepreneurship Development, Bharati Vidyapeeth (Deemed to be University), Pune, Maharashtra
Email Id: Shubhashri.Waghmare@bharatividyapeeth.edu
[3]PhD Scholar, Department of Industrial and Production Engineering, Guru Ghasi Das Viswavidyalaya, Bilaspur, Chhattisgarh, India
Email Id: love.thawait2015@gmail.com
[4]Sr. Lecturer, Department of Elex and Telecommunications, M.H.S.S. College, Mumbai University, Mumbai, Maharashtra, India
Email Id: darakshan.javed.2012@gmail.com
[5]Sr. Solution Architect AI, Department of Artificial Intelligence, IIT Delhi, Hyderabad, Telangana
Email Id: Machinelearning.Venkat@Gmail.Com
ORCID ID: https://orcid.org/0009-0000-9966-6125

Abstract

The industrial safety systems of today use Artificial Intelligence (AI) to provide advanced functions that enable users to forecast outcomes and make decisions without human intervention. The opaque characteristics of intricate AI models present considerable hurdles in safety-critical contexts where transparency, reliability, and accountability are paramount. The research develops an academic framework which uses Explainable Artificial Intelligence (XAI) to support Intelligent Safety Instrumented Systems (SIS) in creating understandable yet dependable safety assessments. The system establishes operational dependability through its combination of AI-based decision processes with three diagnostic levels and explanation tools and human-operated validation methods. The proposed safety decision-making process allows operators to understand better and meets regulatory requirements while building their trust through explainability which extends across all decisionmaking steps. The research demonstrates how XAI technology transforms traditional Safety Instrumented Systems into modern industrial systems which provide transparency and better user experience.

Keywords: Explainable Artificial Intelligence (XAI); Safety Instrumented Systems (SIS); Transparent Safety Decisions; Industrial Safety

INTRODUCTION

A Safety Instrumented System (SIS) safeguards operations by executing designated control functions to stop dangerous processes during situations that reach unacceptable levels. The operational independence of safety instrumented systems from all control systems which manage the same equipment is essential to maintain SIS performance integrity (Mitchell et al., 2017). An SIS comprises same categories of control components, such as sensors, logic solvers, actuators, and other control apparatus, as a Basic Process Control System (BPCS). The entire SIS control system exists to achieve its specific operational functions. An SIS is fundamentally defined by its inclusion of instruments that monitor process variables (such as flow, temperature, and pressure in a processing facility) for

deviations beyond predetermined thresholds (sensors), a logic solver that analyses this data and renders decisions based on the signal characteristics, and final elements that act upon the logic solver's output to ensure the process attains a safe state (Goeritno et al., 2020). All these components must operate effectively for the SIS to execute its SIF. The logic solver may utilise electrical, electronic, or programmable electronic devices, including relays, trip amplifiers, or programmable logic controllers. Support systems, including power, instrument air, and communications, are typically necessary for the operation of Safety Instrumented Systems (SIS). Support systems must be engineered to provide requisite integrity and reliability (Prayudyanto et al., 2022).

Artificial Intelligence (AI) facilitates the creation of advanced autonomous safety-critical systems wherein Machine Learning (ML) algorithms derive optimised and secure solutions. AI can aid human safety engineers in the development of safety-critical systems. Reconciling advanced AI technology with safety engineering methods and standards is a significant barrier that must be resolved before AI can be completely integrated into safety-critical systems (Perez-Cerrolaza et al., 2024).

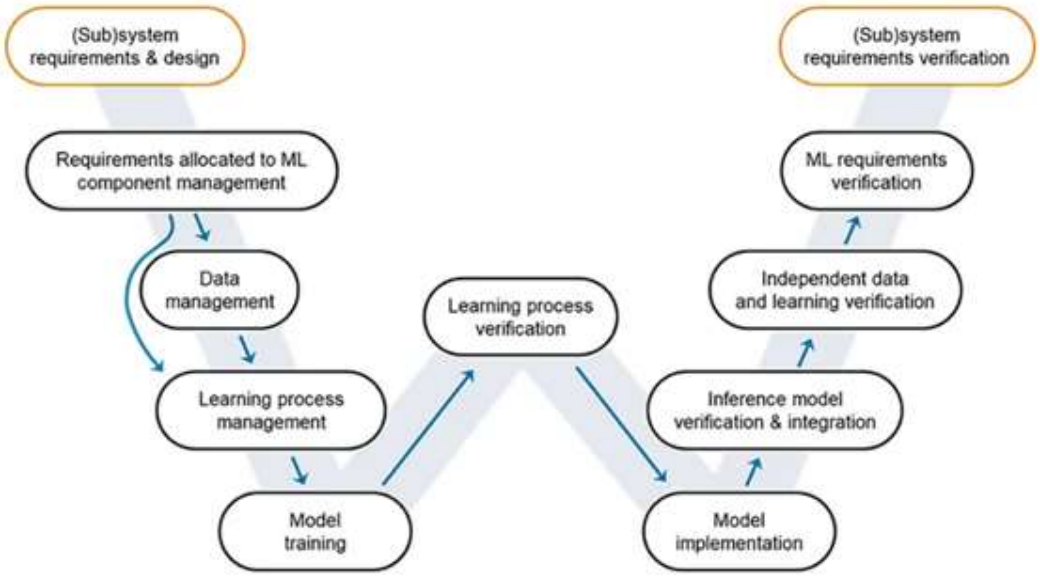

Figure 1. AI in safety critical systems.

SISs which are used in chemical processing and oil and gas and power generation industries must demonstrate dependable performance which operates in a predictable fashion while meeting the requirements of functional safety standards such as IEC 61508 and IEC 61511. The integration of Artificial Intelligence and Machine Learning systems into Safety Instrumented Systems allows for predictive diagnostics and anomaly detection which creates more complex data-driven decision-making systems.

SISs which are used in safety-critical industrial sectors that include chemical processing and oil and gas and power generation need to achieve high levels of reliability together with predictable performance and compliance with functional safety standards which include IEC 61508 and IEC 61511. The use of AI and ML models in SISs enables predictive diagnostics and anomaly detection which makes decision-making processes more complex because they rely on data from multiple sources (Ali et al., 2019).

Most advanced AI models function as "black-box" systems which prevent human operators from understanding their decision-making methods especially deep learning architectures. The absence of system transparency creates multiple safety-critical situations which lead to major operational difficulties because of the following reasons:

1. Decreased operator confidence in automated determinations.
2. Challenges in assessing and verifying system performance.
3. Obstacles in regulatory adherence and safety accreditation.
4. Restricted capacity to do root-cause investigation during failure incidents.

Despite the increasing implementation of AI methodologies in intelligent Safety Instrumented Systems, a significant constraint persists due to the inadequate explainability of model-driven judgements. Traditional SIS designs are constructed upon deterministic logic and rule-based safety functions, which are fundamentally interpretable and verifiable. Conversely, AI-driven SIS depend on intricate, high-dimensional models that possess an absence of intrinsic transparency.

Moreover, current XAI methodologies are typically tailored for general AI applications and fail to sufficiently meet the real-time, high-reliability, and fail-safe demands of SIS systems. This establishes a considerable disparity between AI capabilities and their secure implementation in industrial safety systems. The primary objective of this research is to develop a robust framework that integrates Explainable AI techniques into intelligent Safety Instrumented Systems to ensure transparent, reliable, and interpretable safety decision-making.

This study examines the incorporation of Explainable Artificial Intelligence (XAI) into intelligent Safety Instrumented Systems (SIS) utilised in industrial process control settings. The project uses advanced artificial intelligence and machine learning systems to perform decision-making tasks that require safety assurance through their ability to detect faults and identify anomalies and forecast risks. The research addresses an important challenge which prevents safe implementation of artificial intelligence technology in safety-critical fields because users lack trust in complex AI systems that operate in unknown ways. The proposed method improves safety assurance through its ability to increase transparency and explainability of AI systems which control safety operations, thus enabling users to understand system behavior better and reducing their chances of experiencing dangerous or accidental events. It facilitates regulatory compliance by assuring the traceability and auditability of decisions in accordance with defined safety standards (Akyildiz et al., 2020). Moreover, the incorporation of explainability enhances operational confidence among system operators and stakeholders, while promoting superior decision-making through significant human-machine interaction. This work facilitates the widespread industrial implementation of AI-enabled SISs by offering a comprehensive framework for creating trustworthy, interpretable, and dependable AI systems inside safety-critical industrial contexts.

Understanding Explainable AI

The field of explainable AI has developed because it now evaluates both predictive accuracy and system effectiveness which includes capabilities to understand system operations. XAI fundamentally aims to render the decision-making processes of AI systems transparent and intelligible to human stakeholders. The process involves different elements which include identifying crucial input elements that drove a decision and studying how those elements interact and evaluating whether the model's logic matches established domain knowledge and identifying potential biases and failure points. The XAI field contains various methods which work together as supporting technologies. Model-agnostic methods which include LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) enable users to understand machine learning models through their study of how inputs relate to outputs. LIME produces explanations by altering inputs and analysing the variations in predictions, so producing locally accurate representations of intricate models. SHAP utilises game theory to allocate significance values to attributes, offering local explanations for specific forecasts and global insights into the model's overall behaviour.

Intrinsically interpretable models provide an alternative methodology. The four systems of decision trees, rule-based systems, linear models and Bayesian networks provide complete visibility into their

decision-making processes. The models present a trade-off because they lose some predictive power for accuracy improvements which deep neural networks provide yet their ability to be understood makes them valuable in safety-critical situations where understanding the system takes priority over minor accuracy gains. Current research efforts focus on developing systems which provide built-in interpretability to achieve performance levels that match black-box systems while maintaining complete visibility of their operations.

Attention mechanisms and saliency maps offer visual elucidations, particularly advantageous in computer vision applications prevalent in manufacturing quality control. These strategies elucidate the parts of an image or the time steps in a sequence that most significantly impacted a model's judgement, allowing operators to ascertain that AI systems concentrate on pertinent aspects rather than fallacious associations.

The safety-critical manufacturing field needs explanation methods which can meet the different needs of its stakeholders. The production operators need instant operational information which they can use to verify AI recommendations while solving production issues. The quality assurance teams require complete audit trails which show their compliance with established standards. Engineers need information which helps them solve system problems and improve models while understanding how failures occur. The regulatory auditors need complete documentation which shows that AI systems function according to established technical principles instead of relying on artificial features or biases.

Significance of Explainable AI

1. XAI establishes user trust through its transparent AI system design which enables users to comprehend AI decision-making processes.
2. Many industries demand explanation of AI-based decisions. XAI assists organizations with their legal obligations while it helps them avoid particular legal and ethical challenges.
3. Explainability enables developers to find and fix biases and errors and unpredicted results in AI models which leads to development of more trustworthy and accurate systems.

LITERATURE REVIEW

Artificial Intelligence has advanced recently because researchers discovered that explainability becomes essential for understanding systems used in safety-critical environments. The work of Arrieta et al. (2019) discovered that modern machine learning systems face explainability problems which particularly affect complex systems that use deep neural networks and ensemble methods for their operations. The research developed a complete XAI framework which researchers used to create systems that produce interpretable and fair results while staying accountable across different use cases.

Dwivedi et al. (2022) showed that transparent systems which people can understand need to exist to build trust in AI systems which people use in critical applications. The research study organized XAI methods into different categories while it examined programming frameworks and toolkits which help practitioners choose and use explainable models in actual systems. Holzinger et al. (2022) gave a brief summary of main XAI methods which showed the field had developed enough to explain complex model outcomes through LIME and SHAP and Integrated Gradients. The research of Ali et al. (2023) showed that AI systems need explainability because their black-box design makes it difficult to understand their predictions while the research examined new trends and tools and evaluation methods which established XAI as a rapidly developing research field.

The research of Boudjoghra and Innal (2023) used SISs under IEC 61508 standards to demonstrate industrial safety risk reduction through HAZOP and LOPA risk assessment methods. The study demonstrated that industrial safety depends on three factors which need to be assessed for reliability and Safety Integrity Levels (SIL) and systematic risk assessment needs to determine industrial safety

measures.

Park and Kang (2024) studied how AI functions within Safety 4.0 by demonstrating that IoT and predictive analytics and digital twin technologies enable companies to implement proactive safety management across different sectors. They discovered essential difficulties which included ethical problems and data protection issues and the inability of AI-powered safety systems to offer understandable explanations of their operations. Shah and Mishra (2024) demonstrated how AI transforms occupational health and safety through its capability to conduct predictive risk management and real-time monitoring because these features lead to safer work environments.

The existing research shows that AI-based safety systems require explainability which leads to a strong connection between these two elements. The research field needs to establish better connections between Safety Instrumented Systems and explainable artificial intelligence while current research has made advancements in both XAI and intelligent safety systems.

Current research has developed Handle Explainable Artificial Intelligence and AI-based safety systems into separate research areas that study explainability and SISs as separate domains without studying their application in time-sensitive safety environments. The present XAI methods do not meet industrial safety standards because they fail to provide dependable performance and time-based interpretability and regulatory compliance with IEC 61508 requirements. The research study chooses this topic to create a unified framework which connects explainability with intelligent SISs to enable transparent dependable industrial safety decision-making that meets standards.

METHODOLOGY

This research employs a conceptual and system-integration methodology to design and create a SIS enhanced with XAI. The technique emphasises the integration of explainability, safety protocols, and platform-level implementation into a cohesive and transparent safety decision-making framework.

Development of Explainability Techniques

In this study, Explainable Artificial Intelligence (XAI) experts conceptually collaborate with case-based industrial scenarios to identify and develop effective techniques for interpreting AI model behavior and decision-making processes (Sheu, 2020). These techniques are systematically organized into a modular and reusable library that facilitates the interpretation of model predictions, identification of key influencing features, and generation of human-understandable explanations. The proposed library enables two types of explainability research which include local explainability that delivers individual decision explanations and global explainability which shows complete model operation. The dual system of the explainability framework permits multiple safety-critical applications because it can adapt to different operational needs while increasing system transparency and trustworthiness and user-friendliness of AI-powered SISs.

Integration of Safety Patterns and Architectures

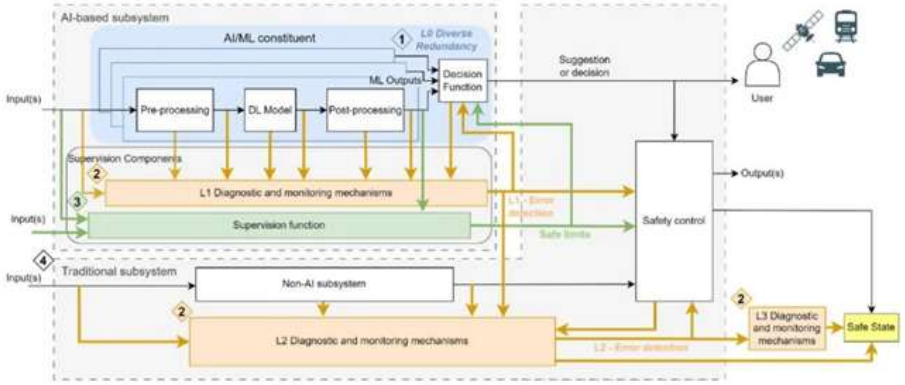

Figure 2. Safety pattern.

The role of safety engineering experts in the proposed system development is highly important because they will specify standardized safety patterns and architectural models that can be used in the Safety Instrumented Systems (SIS). They have made contributions in the design of fail-safe and fault-tolerant system architectures that provide continuous and secure operation even in the presence of abnormal conditions (Buhrmester et al., 2021). Also, the safety patterns are created beforehand to respond and alleviate possible hazards successfully, allowing the organized and uniform safety responses. By utilizing the data from the different test scenarios and reference case studies will be validating the conceptual assessment, which gives a basis of evaluation of system behavior in different risk situations. Taken together, these safety patterns and architectural approaches so that the system satisfies fundamental requirements of reliability, redundancy and regulatory compliance, which are important in safety-critical industrial settings.

Diagnostics and monitoring components

The Level 2 diagnostics and monitoring layer is tasked with a thorough evaluation of system-level health within the suggested framework. It includes the oversight of application nodes running AI models, together with diverse diagnostic and monitoring components, as well as supervisory and decision-making modules. This layer functions by always assessing the performance and interactions of all associated components to identify potential faults, inefficiencies, or inconsistencies. The L2 layer guarantees overall stability, availability, and reliability by adopting a comprehensive perspective of the system's operational status. It is essential for maintaining continuous system operation and facilitating strong, secure, and reliable decision-making in intelligent Safety Instrumented Systems.

Decision-Making and Validation Mechanism

The decision-making and validation method is structured to provide robust, reliable, and trustworthy safety decisions by synthesising outputs from various system components. The decision function fundamentally conducts confidence evaluation by evaluating the dependability of AI-generated outputs through uncertainty estimating techniques. This allows the algorithm to assess the trustworthiness of each prediction and pinpoint instances where judgements may necessitate additional verification.

A redundant model comparison strategy is utilised, wherein numerous AI models execute the same task, such as hazard detection, employing various computational strategies. The outputs of these models are methodically evaluated to identify discrepancies and diminish the probability of incorrect decisions.

Human-in-the-Loop Validation

The system achieves automated operations to a significant extent which requires human operators to ensure operational dependability and system accountability. The team must verify AI outputs by conducting real-time testing of system components. Operators examine XAI module explanations to understand decision-making processes which helps them build trust and develop operational awareness. Human operators possess the power to reject or modify system decisions during periods of uncertainty. This process guarantees that final actions will match expert judgement and safety regulations.

RESULT

The suggested Explainable Artificial Intelligence (XAI) integration in Intelligent Safety Instrumented Systems (SIS) is likely to bring multiple important advantages to safety-related industrial settings. To start with, the framework is expected to enhance the confidence in AI-based safety decisions with the ability to give transparent and understandable reasons underpinning automated behavior, reducing apprehension of black-box models. Secondly, it will increase the transparency of the decisions, which will allow the operators and stakeholders to see the rationale behind every safety response that is essential to validation, auditing, and compliance with regulations.

Moreover, the system will probably enable enhanced human-AI cooperation as operators can freely interact with AI products with meaningful explanations, which result in more informed and assertive decision-making. This synergistic response is what makes sure that human skill does not overshadow automated intelligence particularly during emergencies. Finally, explainability is likely to play a role in minimizing safety risks because explainable outputs would help to spot anomalies early, discover possible failures, and take corrective measures on time. All these effects are indicators of the creation of a safer system that is more reliable, transparent, and human-centered.

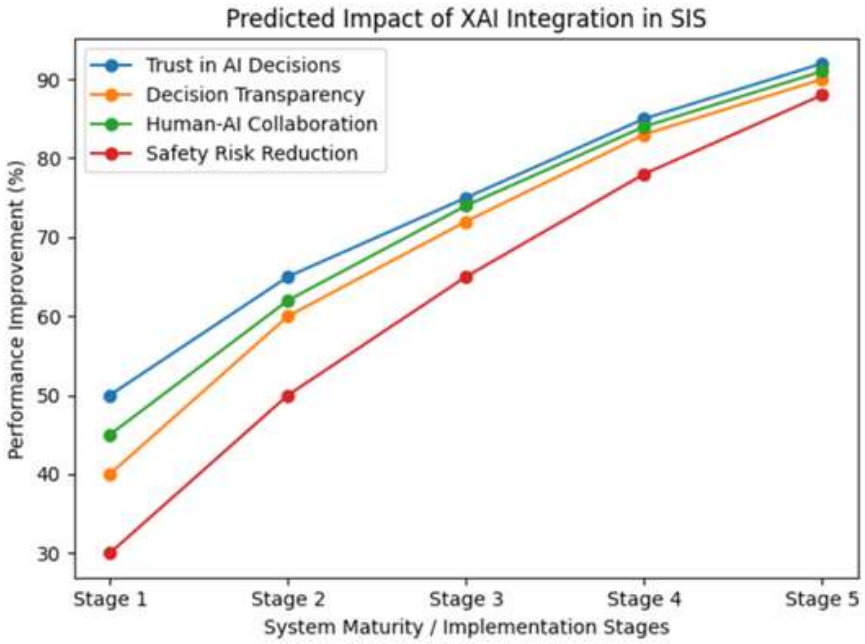

Figure 3. Performance Gains from XAI Integration in Safety Instrumented Systems.

The diagram shows how XAI implementation in SISs establishes better performance results through improved trust and transparency and better human–AI interaction and risk control. The system displays continuous improvement because all indicators show upward growth which demonstrates that explainability helps people understand AI decisions while it increases safety and reliability and operator confidence in critical industrial environments.

DISCUSSION

The results of the present research underscore the transformative nature of explainability in filling the gap between the developed AI and the high standards of safety-critical systems. Although AI is more effective in enabling the Safety Instrumented Systems to handle complex and high-dimensional data, its interpretability has historically impeded its practical use in industrial safety. The suggested framework will overcome this shortcoming by integrating explainability into the decision-making pipeline directly, which will allow a more transparent and responsible system.

An important lesson of the conceptual assessment is that explainability does not just enhance system interpretability but also enhance system robustness in general. The combination of multi-level diagnostics, such as data-level and system-level monitoring, is part of the reason behind better fault detection and system stability. Moreover, the redundancy of AI models and confidence-based decision-making mechanisms minimizes the probability of harmful or unsafe behavior, increasing the reliability of the system.

The other critical factor is the human operators involved in the loop. Instead of eliminating human expertise, the framework helps to facilitate a collaborative method in which human judgment supplements AI-driven knowledge. This especially comes in handy in critical situations where a

contextual knowledge and experience are required to make the final decision. Also, the explainability layer will provide more efficient communication between the operators and the system to provide a better situational awareness and become more responsive.

Although these benefits exist, some challenges abound. The application of XAI methods in real-time industrial systems must be done with a close focus on computational efficiency and latency. Besides, the standardization of the explainability technique to meet the safety standards remains a developing field. All these factors indicate that more research is necessary to streamline and confirm the proposed framework in practical settings.

CONCLUSION

The research presents an original framework which demonstrates how Explainable Artificial Intelligence can be integrated into Intelligent SISs to create transparent safety decision-making processes. The proposed solution addresses major challenges which arise from the lack of transparency found in conventional AI systems through its combination of AI-driven analytics and explainability tools and diagnostic capabilities and human monitoring elements. The framework improves system reliability while increasing operator trust and ensuring safety decisions remain understandable and actionable. The study demonstrates that XAI implementation will significantly improve both the effectiveness and adoption rate of industrial AI-based safety systems.

Future Scope

The proposed framework has numerous opportunities for subsequent research and improvement. The conceptual model can be adapted for real-time industrial application utilising live statistics from industries such oil and gas, manufacturing, and power systems. Secondly, subsequent research may investigate the use of sophisticated explainability methodologies, encompassing hybrid and domain-specific XAI models, to enhance interpretability and precision.

The integration of digital twins with IoT monitoring systems creates a pathway which enables real-time simulation and predictive safety assessment. The initiatives will establish standard evaluation metrics and regulatory frameworks which explainable AI systems must follow to achieve safety standards required in safety-critical environments. The research field requires more research to improve XAI algorithms which need to function in resource-limited environments.

REFERENCES

1. Mitchell KJ, Herena P, Longendelpher TM, Kuhn MC. Kenexis Safety Instrumented Systems Engineering Handbook. Columbus, OH: Kenexis Consulting Corporation 2017.
2. Goeritno A, Nurmansyah D, and Maswan M. Safety Instrumented Systems to Investigate the System of Instrumentation and Process Control on the Steam Purification System. International Journal of Safety and Security Engineering (IJSSE), 2020;10(5):609-616,
3. Prayudyanto MN, Goeritno A, Al Ikhsan SH, and Taqwa FML. Designing a Model of the Early Warning System on the Road Curvature to Prevent the Traffic Accidents. International Journal of Safety and Security Engineering (IJSSE) 2022; 12(3):291-298.
4. Jon Perez-Cerrolaza, Jaume Abella, Markus Borg, Carlo Donzella, Jesús Cerquides, Francisco J. Cazorla, Cristofer Englund, Markus Tauber, George Nikolakopoulos, and Jose Luis Flores. 2024. Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey. ACM Comput. Surv. 56, 7, Article 176 (July 2024), 40 pages.
5. Buhrmester, V.; Münch, D.; Arens, M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Mach. Learn. Knowl. Extr.* 2021, *3*, 966–989.
6. Sheu, Y.h. Illuminating the black box: Interpreting deep neural network models for psychiatric research. *Front. Psychiatry* 2020, *11*, 551299.

7. Akyildiz, Ian F., Ahan Kak, and Shuai Nie. "6G and Beyond: The Future of Wireless Communications Systems." IEEE Access 8 (January 1, 2020): 133995–30.

8. Ali, Muhammad Salek, Massimo Vecchio, Miguel Pincheira, Koustabh Dolui, Fabio Antonelli, and Mubashir Husain Rehmani. "Applications of Blockchains in the Internet of Things: A Comprehensive Survey." IEEE Communications Surveys & Tutorials 21, no. 2 (January 1, 2019): 1676–1717.

9. Chen, S., Liu, J., Chen, C., Xie, S., & Cheng, Z. (2024). Hybrid Explainable Network Intrusion Detection Framework Based on Shapley Additive Explanations. 2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA), 1197–1202.

10. Blesch, K., Wright, M. N., & Watson, D. (2023). Unfooling SHAP and SAGE: Knockoff Imputation for Shapley Values. Explainable Artificial Intelligence, 131–146.

11. Boudjoghra, N., & Innal, F. (2023). Evaluation of Safety Instrumented System in a Natural Gas Facility According to IEC 61508 Standard. *International Journal of Safety and Security Engineering*.

12. Park, J., & Kang, D. (2024). Artificial Intelligence and Smart Technologies in Safety Management: A Comprehensive Analysis Across Multiple Industries. *Applied Sciences*.

13. Shah, I., & Mishra, S. (2024). Artificial intelligence in advancing occupational health and safety: an encapsulation of developments. *Journal of Occupational Health*, 66.

14. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Rana, O., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2022). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, 55, 1 - 33.

15. Arrieta, A., Rodríguez, N., Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion*, 58, 82-115.

16. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J., Confalonieri, R., Guidotti, R., Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion*, 99, 101805.

17. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022). Explainable AI Methods - A Brief Overview. , 13-38.