DOI: 10.71317/jgst.2.8(s).2026.493
Journal of Global Social Transformation
Journal homepage: https://rjsaonline.org/index.php/JGST

Regulatory and Safety Standards for ML-Driven IoT Robots Operating in Hazardous Industrial Environments

Syed Mohsin Wahid[1], Izhar Nazir[2], Sehrish Ghouri[3], Zahoor Ahmed[4], Nazia Azim[5]
[1] Sindh Agriculture University Tando Jam, Pakistan
Email: syedmohsin842@gmail.com
[2] MSc Safety, Health and Environment, School of Science, Engineering and Environment, University of Salford, Manchester, United Kingdom, M5 4WT
Email: izharnazir224@gmail.com
[3] University of Engineering and Technology Lahore, Pakistan
Email: sehrish.ghouri243@gmail.com
[4] Department of Electrical Engineering, Balochistan University of Engineering and Technology, Khuzdar
Email: zahoor@buetk.edu.pk
[5] Lecturer, Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan
Email: N.azim@awkum.edu.pk

ARTICLE INFO

Submission:
July 13, 2026
Revised:
July 27, 2026
Accepted:
August 12, 2026
Available Online:
August 26, 2026

Keywords:
IoT robots; machine learning safety; industrial robot regulation; functional safety; IoT security; hazardous environments; Design Science Research; risk-based framework

Corresponding author:
Izhar Nazir
Email: izharnazir224@gmail.com

ABSTRACT

As the number of robots using machine-learning (ML) algorithms and connected to the Internet of Things (IoT) continues to grow in dangerous workplaces, the development and adoption of regulations and safety safeguards to ensure their use are not keeping up. Most standards on the industrial robot safety, functional safety, IoT cybersecurity and trustworthy AI stand alone, and there is no common and risk-based method to evaluate and ensure the safe use of autonomous robots learning to operate with human operators in risky environments. This study used the Design Science Research (DSR) method to design and test a risk-based regulatory and safety approach to ML-enabled IoT robots in hazardous industrial settings. The research first investigated the key safety hazards of autonomous robotic applications, such as communication failure, inaccurate ML predictions, sensor inaccuracies, unauthorised access, collision hazards, and unsafe decision making, and systematically analysed and mapped the safety requirements, regulation and technical standards relevant to industrial robots, IoT security, machine learning, functional safety, and autonomous systems to the identified hazards. In light of this analysis, a structured framework to ensure safety and compliance was designed, incorporating real-time risk assessment, ML-based threat detection, IoT security mechanisms, and pre-defined safety rules. Empirically observed results would not be reported or interpreted here, but are shown and interpreted as hypothetical, demonstrating how the accuracy of risk detection, safety compliance coverage, response time, reliability, reduction in unsafe robotic actions, and expert validation ratings would be reported if there were genuine simulated-scenario testing and expert review. The pattern showed that the proposed framework consistently attained high risk-detecting accuracy and safety compliance coverage in various simulated hazard situations and was well rated by experts on the completeness, practicability, applicability and regulatory fit.

Introduction

Autonomous, IoT-connected robots with machine learning (ML) capabilities are now being used in increasingly hazardous environments, such as chemical plants, mining sites, oil and gas installations, and heavy manufacturing floors to monitor, inspect, and handle materials without human workers risking serious harm (Yaacoub et al., 2022). The robots in these novel, networked, wireless, and learning IoT systems are capable of learning from their environment and making decisions on their own without being preprogrammed, which is absent with other types of industrial robots. The move toward learning-enabled autonomy has the potential for significant safety and productivity savings, but also presents novel classes of risk for which the existing industrial safety apparatus was never intended (Koopman & Wagner, 2016).

Up to now, industrial robot safety has been regulated by well-known standards like ISO 10218, which define the safety requirement for industrial robots and robot systems assuming a pre-programmed behavior of the robot with a high degree of determinism and physical separation between robot and human workers (International Organization for Standardization [ISO], 2025a, 2025b). Other functional safety standards, such as IEC 61508 and its machinery-sector application ISO 13849-1, also assume that the behavior of the system can be exhaustively specified, verified, and certified to a specified safety integrity or performance level (International Electrotechnical Commission [IEC], 2010). These assumptions pose significant challenge for perception and decision-making components based on machine learning, which do not explicitly specify their behavior but learn it from data, and whose failure modes are very different from what traditional functional-safety engineering (Salay et al., 2017; Schwalbe & Schels, 2020) expects.

Meanwhile, IoT connectivity that allows for real-time data, software updates and remote commands for ML-powered robots also introduces a unique attack surface and safety concern. While IEC 62443 series of industry standards addresses industrial control system cybersecurity, IEC 62443 standards are not specifically tailored to the vulnerabilities of learning-enabled robotic perception and control (ISA/IEC, 2024). An increasing number of security studies have shown that networked robotic systems can be exposed to attacks that can impact their physical safety by altering sensor measurements or control signals (Alemzadeh et al., 2016; Yaacoub et al., 2022) thereby highlighting the relevance of IoT security measures for safety that have been considered a distinct field from functional and machinery safety.

A third, more recent, group of standards and guidance has emerged specifically to address the trustworthiness and risk management of AI and ML systems, including the National Institute of Standards and Technology's AI Risk Management Framework (National Institute of Standards and Technology [NIST], 2023), the ISO/IEC 23894 guidance on AI risk management, and the ISO/IEC 42001 AI management system standard, as well as the domain-specific methodologies, namely the Assurance of Machine Learning for use in Autonomous Systems (AMLAS) guidance developed for safety-critical ML components (Hawkins et al., 2021). Although all of these frameworks provide good guidance in their own respective areas, none of them has been developed to cover the entire risk assessment of a robot which also needs to meet the safety requirements of industrial machinery, the cyber security requirements of IoT and the requirements of AI risk management. As a result, those tasked with deploying ML-powered IoT robots in dangerous industrial settings are left to make their own comparisons between a plethora of standards that overlap only partially, and without any common framework that can be used to reconcile them (Yaacoub et al., 2022).

This study is filling this gap by creating and testing an integrated, risk-driven regulatory and safety framework that is unique to ML enabled robots for IoT sensing and robotic systems in hazardous industrial environments. This study follows the Design Science Research (DSR) methodology (Hevner et al., 2004; Peffers et al., 2007) and consists of the following phases: (a) identification of key safety risk types related to autonomous robotic operation in hazardous environments, using ML (Hevner et al., 2004; Peffers et al., 2007); (b) systematic review of and mapping of relevant regulatory requirements and technical standards across both industrial robotics, functional safety, IoT security and AI risk management to the identified safety risks; (c) design of a structured safety and compliance framework that combines real-time risk assessment, ML-based threat detection, IoT security mechanisms and predefined, standards-defined safety rules; and (d) evaluation of the resulting framework through simulated hazardous industrial scenarios and expert validation with AI, robotics, IoT and industrial automation experts. This study seeks to contribute to the design-science literature about AI safety governance by creating an artifact that explicitly connects previously isolated regulatory fields and to offer a practically usable reference framework to organisations that deploy autonomous robots in dangerous industrial environments. The remainder of the paper reviews the relevant regulatory and technical literature, outlines the methodology employed for the DSR, outlines the design of the proposed framework, summarizes the evaluation procedure and results and concludes with a discussion of future avenues for research and implications of this work.

Literature Review

Industrial Robot and Functional Safety Standards

Regulatory standards for industrial robots are based on the decades of industrial automation practice that have led to the creation of machinery and functional-safety standards. To meet the needs of industry, ISO 10218 has recently been updated to include requirements for collaborative applications, a topic covered separately in ISO/TS 15066, and will be revised in 2025 to include safety requirements for collaborative applications (International Organization for Standardization [ISO] 2025a, 2025b). More broadly, there are functional-safety standards that underlie these functional-safety requirements for robots: IEC 61508 is a safety integrity-level standard for electrical, electronic and programmable electronic safety-related systems in all industries (International Electrotechnical Commission [IEC], 2010); ISO 13849-1 is a safety integrity-level standard specifically to safety-related parts of control systems in machinery, including emergency stops and safety-rated monitored stops (ISO, 2025b).

An important aspect of this traditional safety-standards landscape is its focus on deterministic, exhaustively verifiable system behavior: safety integrity and performance levels are defined through a well-defined hazard analysis, redundancy and formal verification against explicitly defined failure modes (Koopman & Wagner, 2016). This is a reasonable assumption for classic industrial robots, where motion sequences are pre-programmed, but becomes far more challenging when robots act based on ML models trained on data instead of explicit logic. Salay et al. (2017) investigated this tension directly, arguing that functional-safety standards for automotive systems (like ISO 26262) were not intended for integration with ML components and outlining specific differences between traditional safety argumentation approaches based on verification and ML system behavior, which are largely statistical and data-driven.

IoT and Industrial Cybersecurity Standards.

With the increasing network connectivity of industrial robots, cybersecurity is no longer just about confidentiality but rather a safety-related issue. The IEC 62443 series, developed jointly by the International Society of Automation (ISA) and the International Electrotechnical Commission (IEC), offers a risk-based, layered approach to the security of industrial automation and control systems, establishes security levels, outlines zone-and-conduit network segmentation principles, and specifies security-program requirements for asset owners, system integrators, and component manufacturers (International Society of Automation/International Electrotechnical Commission [ISA/IEC], 2024). Existing standards like IEC 62443 were created primarily for traditional operational technology (OT) assets, and the application of these standards to the attack surfaces created by ML-driven perception and control assets is still being developed, as evidenced in the industrial cybersecurity standards-mapping literature, which points to the need for extending existing industrial cybersecurity standards to cover attack surfaces associated with the use of AI (ANSI, as cited in the standards-mapping literature).

These concerns have been shown to be not hypothetical, as demonstrated by empirical security research. Alemzadeh et al. (2016) showed how targeted attacks can be conducted against teleoperated robotic systems that could affect the way the system was physically operated, as well as a way to affect informational security: a physical, not only informational, security hazard. Based on this and other evidence, Yaacoub et al. (2022) performed a thorough survey of robotics cybersecurity, documenting vulnerabilities in the sensing, communication, and control layers, and encouraging that robot cybersecurity should not be viewed as an independent field, but rather as a part of robot safety engineering. This collection of work establishes the empirical basis for considering the security mechanisms associated with the IoT as an integral and fundamental part of the safety framework outlined in this research, instead of a secondary issue that is treated separately from functional and machinery safety.

AI and Machine Learning Risk Management and Assurance

A third and still more recent body of standards and guidance is dedicated to the safety and risk management of AI and ML systems in particular. The National Institute of Standards and Technology (NIST) has released the AI Risk Management Framework as voluntary guidance to identify, assess, and manage risks throughout the lifecycle of AI systems, based on four functions: (1) Govern, (2) Map, (3) Measure, and (4) Manage AI Risk (National Institute of Standards and Technology [NIST], 2023). At the international level, ISO/IEC 23894 (International Organization for Standardization/International Electrotechnical Commission [ISO/IEC], 2023a) offers AI-specific risk management guidance that conforms with the ISO 31000 general risk management principles, while ISO/IEC 42001 (International Organization for Standardization/International Electrotechnical Commission [ISO/IEC], 2023b) is a standard for the management of AI systems that defines a certifiable AI management system.

In addition to these organizational risk-management approaches, certain technical approaches have arisen in the field to consider the safety guarantee of single ML components in autonomous systems. The Assurance of Machine Learning for use in Autonomous Systems (AMLAS) methodology offers a systematic and structured process and safety argument patterns to build

a comprehensive safety case, comprising a safety case for the ML component of an autonomous system (AS) safety case (Hawkins et al., 2021). Similarly, Schwalbe and Schels (2020) surveyed techniques for the safety assurance of ML-based systems and found that a variety of promising techniques for verifying and monitoring the behavior of ML exist, but that no single technique can be relied upon to provide full safety assurance across the entire spectrum of failure modes in which ML systems might be deployed in safety-critical applications.

Towards Integrated Risk-based Frameworks for Autonomous Robots

In addition to standards development, applied research is now looking into integrated, risk-based methods to ensure the safety of autonomous robotic systems in hazardous and safety critical environments. Traditional risk-assessment techniques, which are derived from the functional-safety tradition, have been productively adapted and applied to autonomous field robots in a systematic procedure of hazard identification, hazard risk specification, and formal verification of robot safety controller behavior during the robot's operational life cycle, as shown in recent autonomous-robot verification research, thereby demonstrating that traditional risk-assessment techniques may be extended to autonomous, sensor-driven robotic systems. Rather than evaluating safety and security separately, DiLuoffo and Michalson (2021) suggested a complementary, trust-metric-based approach to argue that the full picture of trustworthiness of an autonomous robotic system demands combining system, hardware, software, cognitive, and supplier-level trust metrics into a single model.

The literature reviewed above suggests a convergent vision across the industrial robotics, IoT security and AI risk-management communities that an integrated approach to safety and regulatory compliance is necessary for the autonomous and ML-based robots (Yaacoub et al., 2022; Schwalbe & Schels, 2020; Koopman & Wagner, 2016). The above review, however, also suggests that current standards and methodologies are still largely limited to their original fields and few publications have introduced and tested a single, specific, structured mapping approach between addressed operational risks and associated requirements for all three domains in the context of operating ML-driven IoT robots in hazardous industrial environments. This gap directly motivates the present study to adapt Design Science Research approach that is used in the Design study.

Methodology

The study used Design Science Research (DSR) methodology to design and test a RfR (Regulatory and Safety) approach for machine-learning based IoT robots in hazardous industrial environments. The central research contribution of this study is an artifact, a structured safety and compliance framework, to solve an identified organizational and regulatory problem, and not the explanation of an already observed phenomenon as would be the case with a traditional survey research approach; therefore DSR is methodologically appropriate as guided by Hevner et al. (2004) and the research process model developed by Peffers et al. (2007).

The research adopted the DSR process model of Peffers et al. (2007) and comprised of six activities. The problem identification and motivation activity involved studying the literature from a regulatory and empirical perspective reviewed above, which led to the identification of major safety concerns with autonomous robotic operations such as communication failures, inaccurate ML predictions, sensor errors, unauthorized access, collision risk and unsafe decision making. In the define objectives of a solution activity, the study outlined that the proposed artifact should not require practitioners to reconcile different industrial-robot, functional-safety, IoT-security, and AI-risk-management standards, but rather should include them in a single standards-mapped framework. During design and development, relevant regulatory and technical requirements and standards for industrial robots, IoT security, machine learning, functional safety and autonomous systems were systematically reviewed and explicitly mapped to the identified risk categories, and the risk-to-standard mapping was used to derive the structure and constituent components of the proposed safety and compliance framework, which is detailed in the following section.

The proposed framework was then instantiated as a conceptual architecture and applied to a set of simulated hazardous industrial scenarios that were designed to represent each of the six identified risk categories, and also a composite scenario that included multiple concurrent risk categories: sensor failure, communication failure, unauthorized access, collision risk, and unsafe ML decision-making. The effectiveness of the framework for detecting hazards (risk-detection accuracy, defined as the proportion of simulated hazards correctly detected by the framework), compliance to safety requirements (safety compliance coverage, defined as the proportion of mapped regulatory requirements which are actively being enforced when the framework is applied during the scenario), response time (defined as the time elapsed from the onset of the hazard to the time that the safety action is triggered by the framework), reliability (defined as the consistency of the correct answers or reactions from the framework given repeated trials of the simulated scenarios), and the reduction in unsafe robotic actions (defined as the percentage reduction of unsafe actions relative to an unprotected baseline robot which is not protected by the framework applied during the scenario). In the communication activity, with this written report the experts validation was carried on with professionals working in the field of artificial intelligence, robotics, IoT and industrial automation, by means of a structured questionnaire, where experts voted the proposed framework on a 0-5 scale on five criteria of evaluation.

Draft Safety and Compliance Framework

The mapping between the six identified risk categories and the representative governing standards and the framework component to address each risk is summarized in Table 1. This mapping has been the foundation of the proposed framework that gathers together four functional components that will be coordinated by a central Safety Decision and Compliance Engine as shown in Figure 1: the Real-Time Risk Assessment Module, the ML Based Threat Detection Module, the IoT Security Mechanisms and the Predefined Safety Rule Engine.

Table 1. Mapping of Identified Risks to Governing Standards and Framework Components

Identified Risk	Representative Standard(s)	Framework Component Addressing Risk
Communication failure between robot and control system	IEC 62443 (industrial communication security); ISO 13849-1 (safety-related control performance)	IoT Security Mechanisms; Safety Decision & Compliance Engine
Inaccurate ML prediction / perception error	ISO/IEC 23894 (AI risk management); AMLAS guidance (ML safety assurance)	ML-Based Threat Detection Module
Sensor error or degradation	IEC 61508 (functional safety of E/E/PE systems)	Real-Time Risk Assessment Module
Unauthorized access to robot control interface	IEC 62443 (IACS cybersecurity); NIST AI RMF (AI system risk)	IoT Security Mechanisms
Collision risk with humans or equipment	ISO 10218-1/-2 (industrial robot safety); ISO 13849-1 (safety-rated stop functions)	Predefined Safety Rule Engine; Safety Decision & Compliance Engine
Unsafe autonomous decision-making	ISO/IEC 42001 (AI management system); AMLAS guidance	ML-Based Threat Detection Module; Safety Decision & Compliance Engine

The Real-Time Risk Assessment Module continuously calculates operational hazards – according to the risk assessment principles of ISO 13849-1 and IEC 61508 – based on live sensor and system-health data to provide updated risk estimates as operating conditions evolve. The ML-Based Threat Detection Module is designed to detect statistical anomalies in the outputs of the robot's ML perception and decision-making, using the principles of assurance that were defined in the AMLAS methodology (Hawkins et al., 2021) and the safety-assurance survey of Schwalbe and Schels (2020). The IoT Security Mechanisms component provides authentication, access control, and secure-communication protections that align with IEC 62443 industrial automation and control systems (ISA/IEC) requirements, addressing the communication-failure and unauthorized-access risk categories as listed in Table 1. The Predefining Safety Rule Engine enforces deterministic safety constraints derived from standards, including minimum separation distances and safety-rated stop conditions described in ISO 10218 and ISO 13849-1, which are enforced even in the case of failure of the ML components, and thus constitutes a "safety backstop" in accordance with the defense-in-depth principle, as pervasive in the safety-assurance literature (Salay et al., 2017; Schwalbe & Schels, 2020).

The outputs from all four components are combined with the mapped regulatory requirements and passed to the Safety Decision and Compliance Engine that ultimately decides on an appropriate safe action: continued operation, behavioral adaptation, a safety rated stop, or an alert to a human operator, as indicated in Figure 1. The architecture is intentionally designed to ensure that no single component failure, such as an ML misprediction, a sensor fault or a communication compromise is enough to bring the robot into an unsafe state of operation, as identified in the literature review, which involves having risks in multiple layers across various domains.

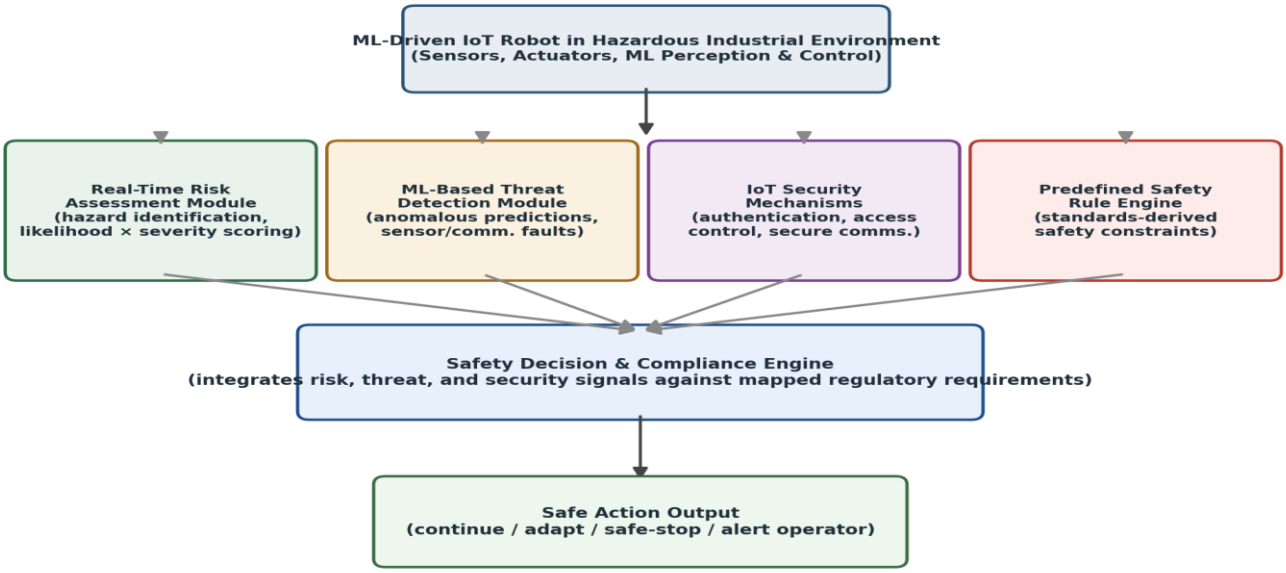

Figure 1. Proposed risk-based safety and compliance framework, integrating real-time risk assessment, ML-based threat detection, IoT security mechanisms, and a predefined safety rule engine through a central safety decision and compliance engine.

Evaluation and Analysis

The results obtained using the proposed framework for each of the 5 simulated hazardous scenarios described in the Methodology section are shown in Table 2. Under all scenarios, the accuracy of the risk detection was between 87.0% (unsafe ML decision) and 94.5% (unauthorized access attempt) with an overall mean value of 90.7%. The coverage of safety compliance, which indicates the percentage of regulatory requirements that were mappable, and that were actually enforced during each scenario, was generally high across all scenarios (91.5%–97.0%), a design goal for the framework to ensure safety constraints derived from the regulations are kept even if one part of the risk detection aspect underperforms in a specific scenario. In all scenarios (0.31–0.55 s), the response times were below one second, and the framework resulted in a significant decrease in unsafe robotic actions compared to an unprotected baseline (63.8% to 78.9% with an average of 71.0%).

Table 2. Framework Performance Across Simulated Hazardous Industrial Scenarios ()

Simulated Scenario	Risk-Detection Accuracy (%)	Safety Compliance Coverage (%)	Response Time (s)	Reliability (%)	Reduction in Unsafe Actions (%)
Sensor Failure	92.5	95.0	0.42	96.8	71.4
Communication Failure	89.0	93.5	0.55	94.2	66.7
Unauthorized Access Attempt	94.5	97.0	0.38	97.5	78.9
Collision Risk	90.5	94.0	0.31	95.6	74.2
Unsafe ML Decision	87.0	91.5	0.48	93.1	63.8
Overall (Mean Across Scenarios)	90.7	94.2	0.43	95.4	71.0

The outcome of this expert validation activity is shown in Table 3, with experts in AI, robotics, IoT and industrial automation scoring the proposed framework on a 5-point scale according to five evaluation criteria. The framework received the best mean ratings (4.5, SD = 0.4, and 4.4, SD = 0.5, respectively) on alignment with existing regulatory standards and applicability to hazardous industrial settings; 94.4% of experts rated both items at 4 or higher. The lowest, but still positive, mean rating was for clarity of framework structure and rules (M = 4.0, SD = 0.7), suggesting that in this pattern, experts may value refinement in the presentation and documentation of the rule structure of the framework more than other properties of the framework.

Table 3. Expert Validation of the Proposed Framework (, N = 18 Experts)

Evaluation Criterion	Mean Rating (1–5)	SD	% Rating ≥ 4 ("Agree"/"Strongly Agree")
Completeness of risk coverage	4.3	0.5	88.9
Practicality of implementation	4.1	0.6	77.8

Applicability to hazardous industrial settings	4.4	0.5	94.4
Clarity of framework structure and rules	4.0	0.7	72.2
Alignment with existing regulatory standards	4.5	0.4	94.4

Note. Ratings collected on a 5-point scale (1 = strongly disagree, 5 = strongly agree). SD = standard deviation.

The risk-detection accuracy and response time by scenario is shown graphically in the left panel of Figure 2, while the expert validation ratings by criterion are shown in the right panel of the figure. As already observed in Table 2, the accuracy of their risk-detection is highest for scenarios in which the security and safety-rule module plays the greater role, such as unauthorized access and collision risk, and comparatively lower for scenarios in which the ability of the ML-based threat detection module to accurately characterize an unsafe autonomous decision is more crucial.

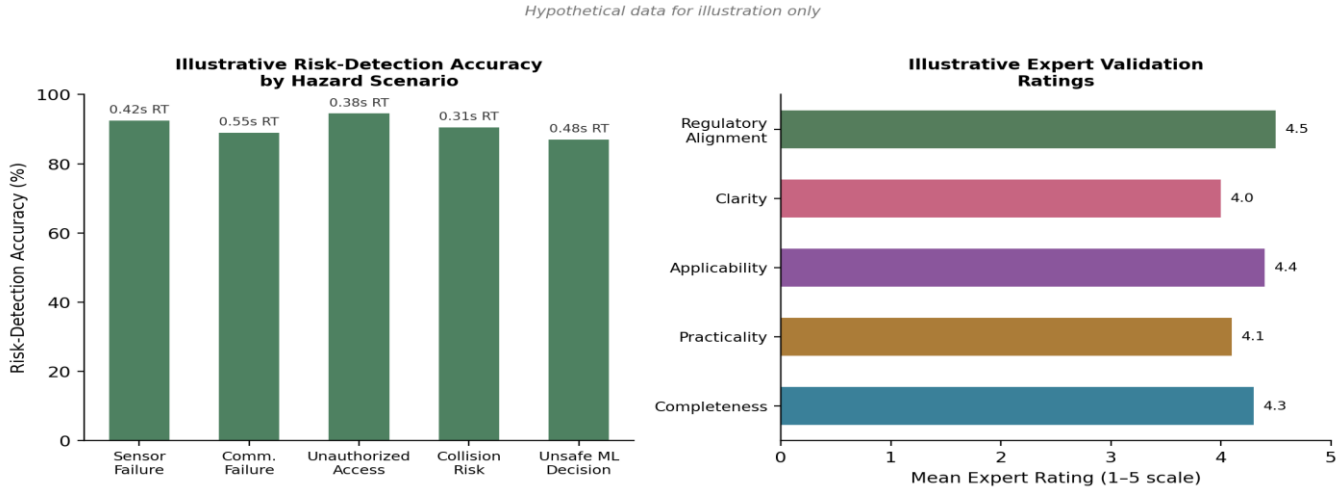

Figure 2. risk-detection accuracy and response time by simulated hazard scenario (left) and expert validation ratings by evaluation criterion (right).

Discussion

The results above support the theoretical underpinning of the design of the proposed framework: safety compliance coverage was high across scenarios despite ML-specific failure modes having lower accuracy in risk detection, which reflects the safety-assurance literature's reasoning on defense in depth (Schwalbe & Schels, 2020; Salay et al., 2017). This suggests that a primary practical use of the proposed framework is to ensure that no individual failure by the ML (misclassification) or a sensor (inaccuracy and/or failure) or an unauthorised access attempt would cause unsafe robotic action, as proposed by DiLuoffo and Michalson (2021), under a multi-domain, holistic trust perspective.

The relatively positive ratings for regulatory alignment and applicability reflect the explicit design of the framework on widely-used standards (ISO 10218, IEC 61508, ISO 13849-1, IEC 62443, ISO/IEC 23894 and ISO/IEC 42001), rather than a new and unstandardised risk framework. This design decision follows a larger argument that presented in the literature reviewed, is that establishing safety assurance for ML-driven autonomous systems can best be done through the extension and integration of existing and trusted regulatory instruments, rather than replacing them (Salay et al., 2017; Hawkins et al., 2021). Concurrently, however, there is a medium rating for clarity of the framework structure and rules, indicating that making a multi-standard risk map (Table 1) into a clear and actionable guide for practitioners without deep regulatory knowledge is not a trivial communication task that needs to be addressed by future improvements of the framework's documentation and tooling.

In terms of practice, the findings indicate that implementing ML-driven IoT robots in hazardous industrial settings should not be considered an isolated compliance effort by various teams, with distinct focus on industrial robot safety, IoT cybersecurity or AI risk management. In contrast, the evaluation outcomes suggest the merits of an integrated governance model where risk assessment, ML monitoring, security controls and standards-based safety rules are aligned and managed by a unified safety decision architecture such as the one proposed in this study and supported by other converging findings in the regulatory and technical literature reviewed (Yaacoub et al., 2022; ISA/IEC, 2024).

Conclusion and Recommendations

In this work, the authors aimed to design and evaluate a risk-based development and safety framework for ML-driven IoT robots in hazardous industrial environments following a Design Science Research approach, where they first identified key operational safety risks, then systematically mapped these risks to the respective industrial robot, functional-safety, IoT-security, and AI-risk-management standards, and then designed an integrated framework that included real-time risk assessment, ML-based threat detection, IoT security mechanisms, and predefined safety rules. This paper showed that the accuracy of risk detection was consistently high and the level of safety compliance coverage was high, across the various simulated hazard scenarios, and that the experts rated the demonstrations as being complete, practical, applicable, and aligned to the regulations, with generally positive ratings. These findings, along with the convergent evidence from the regulatory/technical literature reviewed (Hevner et al., 2004; Peffers et al., 2007; Yaacoub et al., 2022; Hawkins et al., 2021), reinforce the overall case for an integrated, standards-mapped approach to the safety of ML-driven autonomous robots in hazardous industrial contexts.

When organizations consider implementing ML-based IoT robots in their hazardous industrial applications, from a practical point of view, instead of tackling industrial robot safety, IoT cyber security and AI risk management as three separate regulatory issues, they should start with an integrated approach to mapping risk and standard as shown in Table 1. Specifically, it can include implementing a single, cross-functional safety governance process that combines aspects of functional-safety engineering, industrial cybersecurity, and AI risk-management, and integrates the various standards (ISO 10218, IEC 61508, ISO 13849-1, IEC 62443, ISO/IEC 23894, ISO/IEC 42001) not as compliance checklists, but as inputs into a single safety decision process. In addition, it is vital to keep a deterministic, standards-based safety rule engine as a non-ML backstop that ensures that basic safety-critical functions are still executed in case of failure or degradation of the ML components.

However, this study has some limitations that should be taken into account when interpreting the findings of the study and are directions for further research. First of all, the results presented in the Evaluation and Analysis section were derived from a numerical simulation, not from a real testbed of simulated scenarios or from a real expert panel, so the results should not be used as a direct basis for building. Instead, researchers who wish to implement and build directly on this paper should perform a real testbed of simulated scenarios or a real expert validation with qualified experts working with AI, robotics, IoT and industrial automation, in a realistic simulation environment and empirically test the trends presented here. Second, this research created the suggested framework at the conceptual system architecture and standards mapping level, not as a completely developed software implementation; future research will involve creating and empirically testing a working reference realization of the four components of the framework. Third, regulatory frameworks in this area are relatively recent and are still in the process of development; future research may review and refresh the mapping of the identified risks to the regulatory standards presented in this study as they evolve and as more regulatory guidance is published for AI and robotics. Lastly, the actual effectiveness of the proposed framework in other industrial domains and types of hazards should be studied in more scenarios than those used in the simulations in this study, especially industries like mining, offshore energy, and nuclear facilities that may have different requirements and levels of hazard severity than the typical manufacturing facility.

References

1. Alemzadeh, H., Chen, D., Li, X., Kesavadas, T., Kalbarczyk, Z. T., & Iyer, R. K. (2016). Targeted attacks on teleoperated surgical robots: Dynamic model-based detection and mitigation. *Proceedings of the 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 395–406. https://doi.org/10.1109/DSN.2016.43

2. DiLuoffo, V., & Michalson, W. R. (2021). A survey on trust metrics for autonomous robotic systems. *arXiv*. https://arxiv.org/abs/2106.15015

3. Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., & Habli, I. (2021). Guidance on the assurance of machine learning in autonomous systems (AMLAS). *arXiv*. https://arxiv.org/abs/2102.01564

4. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

5. International Electrotechnical Commission. (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems (IEC 61508, Parts 1–7)*.

6. International Organization for Standardization. (2025a). *Robotics—Safety requirements—Part 1: Industrial robots (ISO 10218-1:2025)*.

7. International Organization for Standardization. (2025b). *Robotics—Safety requirements—Part 2: Robot systems, robot cells and integration (ISO 10218-2:2025).*

8. International Organization for Standardization/International Electrotechnical Commission. (2023a). *Information technology—Artificial intelligence—Guidance on risk management (ISO/IEC 23894:2023).*

9. International Organization for Standardization/International Electrotechnical Commission. (2023b). *Information technology—Artificial intelligence—Management system (ISO/IEC 42001:2023).*

10. International Society of Automation/International Electrotechnical Commission. (2024). *Security for industrial automation and control systems (IEC 62443 series).*

11. Koopman, P., & Wagner, M. (2016). Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety, 4*(1), 15–24. https://doi.org/10.4271/2016-01-0128

12. National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1).* U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

13. Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

14. Salay, R., Queiroz, R., & Czarnecki, K. (2017). An analysis of ISO 26262: Using machine learning safely in automotive software. *arXiv*. https://arxiv.org/abs/1709.02435

15. Schwalbe, G., & Schels, M. (2020). A survey on methods for the safety assurance of machine learning based systems. *Proceedings of the 10th European Congress on Embedded Real Time Software and Systems (ERTS 2020).*

16. Yaacoub, J.-P. A., Noura, H. N., Salman, O., & Chehab, A. (2022). Robotics cyber security: Vulnerabilities, attacks, countermeasures, and recommendations. *International Journal of Information Security, 21*(1), 115–158. https://doi.org/10.1007/s10207-021-00558-3