Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26501-cd

Why compliance is not enough: Evaluating AI Systems in Safety-Critical Contexts

Alexandra Fernandes

Equinor ASA, Norway. E-mail: almm@equinor.com

Arne Jarl Ringstad

Equinor ASA, Norway. E-mail: ajri@equinor.com

Jan Tore Ludvigsen

Equinor ASA, Norway. E-mail: jtl@equinor.com

In 2024 the European Union began implementing the Artificial Intelligence Act (EU AI Act) - the first regulation of its type. This has generated interest and concern across industrial and academic forums. The EU AI Act follows a risk classification approach based on the impact that different Artificial Intelligence (AI) systems can have on health, safety, and fundamental rights. However, this conception of risk primarily reflects societal and human rights concerns, being distinct from the typical risk classification approaches in safety critical contexts, such as the oil and gas industry, which focus on preventing major accidents and controlling associated risks.
This discrepancy on the interpretation of risk can result in mismatches in assessment outcomes. For instance, an AI system that is considered "low risk" under the EU AI Act could be "high risk" when assessed through a major accident prevention framework. In this work we explore this inconsistency and analyse its impact for safety critical industries.
We illustrate the issue presenting examples of AI systems that would need to be further investigated, evaluated, and qualified for deployment, even though they would not entail further assessment according to the EU AI Act requirements. We discuss the implications of these findings for technology qualification processes and human-AI interaction evaluations, highlighting the relevance of other pre-existing regulation for safety critical industries. Additionally, we consider how safety critical sectors might benefit from specific requirements and upcoming standards associated with high-risk AI systems under the EU AI Act. We argue that these requirements and guidelines could be valuable beyond the EU AI Act classification paradigm when integrated into pre-existing industry safety frameworks. We highlight the relevance of nuanced perspectives as well as context aware approaches to AI risk assessment, safeguarding that the new regulatory focus does not overlook recognised safety good practices.

Keywords: Artificial intelligence, safety risk, safety critical systems, human-AI interaction, technology qualification, EU AI Act

1. Background

The discourse around Artificial Intelligence (AI) has been visibly optimistic and there is a vast technology hype, with ambitions to accelerant the integration of AI in society at large, including safety-critical industries. This trend can result in a conflict of goals where the pressure to "move fast" is contradictory to the need to know more and assure safe operations. In this context, governance of AI is a central, pushed by the introduction of the European Artificial Intelligence Act (EU AI Act, European Commission) in 2024. The EU AI Act is the first comprehensive regulatory framework for AI systems and adopts a risk-based approach, classifying AI systems according to their potential impact on health, safety, and fundamental rights.

Within safety-critical industries, including oil and gas, current regulatory frameworks provide robust foundations for managing major accident risk, but do not yet contain AI-specific requirements, not necessarily addressing new challenges with emerging technology (probabilistic systems, lower reliability, adaptive

and black box systems). Current safety methods might be difficult to apply to AI technology and insufficient to assure safe deployment. The current regulatory challenges have led several stakeholders to take the EU AI Act as a mechanism to mitigate the current lack of AI-specific regulations in industry regulations. While this is a crucial milestone, there is a growing tendency (both in society and across industries) to place excessive confidence in this one regulatory instrument. We witness an implicit assumption that compliance with the EU AI Act, supported by its associated harmonised standards (Soler Garrido *et al.*, 2024) will ensure efficient management of the full spectrum of risks associated with AI systems in all domains of application. However, pre-existing frameworks (e.g. Markusen, 2024; NIST, 2023; OECD, 2024) have identified risks in systems that might fall outside the EU AI Act's high-risk definition and still require extensive context-specific evaluation and qualification before deployment.

Without recognising the limitations of the regulation(s) and the need for complementary risk assessments rooted in domain specific safety practices, organisations might be overlooking critical hazards that the new legislation was never designed to address.

1.1 *Safety and compliance*

Safety is the top priority in high-risk domains. Achieving safety requires more than regulatory compliance. Laws, regulations and standards often point towards minimum requirements and fail to capture the full picture of safety and risk in dynamic environments with highly complex socio-technical systems.

An exclusive focus on compliance will be insufficient since the focus on adherence to rules dismisses the behavioural, cultural, contextual, and systemic contributors of risk. For instance, Aderamo and collaborators (2024) has highlighted the relevance of behavioural safety programs in high-risk industries and its contribution to resilience and workforce empowerment that are linked to safety outcomes. From an industry perspective, Human Organizational Performance principles have been adopted (e.g. Cocklin, 2012; Norsk Industri, 2025) based on the notion that although compliance and individual commitment are important, they will not suffice to further improve safety in an organization. There is an understanding in both industry and academia that we require knowledge of all capabilities and conditions that will enable safer practices – we need to know the risk to be able to mitigate it.

A relevant principle in the Norwegian regulatory framework for the petroleum sector is the requirement for continuous improvement within health, safety and environment (HSE) management. The Norwegian Ocean Safety Authority (Havtil) stipulates in §23 of the Management Regulations that organizations must not only comply with applicable regulatory requirements, but also systematically identify processes, activities, and products with potential for enhancement. This includes establishing mechanisms to evaluate the need for improvement measures, implementing them, and subsequently assessing their effectiveness. The intention is to ensure that HSE performance develops dynamically over time, rather than remaining at a minimum regulatory baseline.

In the case of AI systems, thinking beyond compliance is of special relevance as most of the current regulatory frameworks face challenges due to the specific nature of these systems that will require dynamic assessments through their lifecycles with uncertainty linked to the learning capabilities and the self-development of the models while in operation. This challenges both the regulations and the existing techniques for technology qualification.

1.2 *Misaligned risk definitions*

The EU AI Act approach primarily addresses societal and ethical concerns, such as discrimination, privacy, and transparency which

are critical for AI deployment, but are predominantly focused on individual harm.

This interpretation of risk differs from the traditional frameworks used in safety-critical industries, such as petroleum, where risk management focuses on major accident prevention (ISO 31000:2018, IEC 61508:2010; NORSOK Z-013:2024). In such contexts risk is defined according to the likelihood and severity of events that could lead to catastrophic consequences, including loss of life, environmental damage, and asset destruction. Regulatory entities such as Havtil in Norway define risk as "the consequences of an activity with the associated uncertainty" and emphasize systematic hazard identification, barrier management, and technology qualification to ensure safe operations.

We argue that the divergence between these two risk paradigms introduces challenges to the assessment of AI systems in safety critical contexts. For example, an AI system classified as "low risk" under the EU AI Act (such as a predictive maintenance algorithm) may still pose significant risk/hazards if integrated into systems that influence process safety or emergency shutdown functions. This highlights the need for context-aware risk assessment processes and methods (Ringstad et al., 2025).

1.3 *Humans and AI*

As AI becomes progressively present in safety critical contexts, understanding the interplay between humans and AI is essential. Aspects such as transparency, trustworthiness, and cognitive workload have direct implications for safety performance and must be considered alongside technical quality and reliability of the systems (e.g. Endsley, 2023). Regulation such as the EU AI Acts emphasises the relevance of these aspects but simultaneously reinforces the need for humans to be "in control" of the technology, most commonly, in a supervisory role (e.g. Kyriakou & Otterbacher, 2023).

As AI systems develop towards more autonomy, traditional compliance-based approaches may fail to capture new safety risks, reinforcing the argument that compliance alone is not sufficient (although it is necessary) to ensure safety in high-risk domains. Simultaneously, as systems grow more complex, intelligent, and adaptive, there is a need to explore and define the technical, human, and interaction challenges in a way that will enable the design of effective mitigation strategies (e.g. Skraaning & Jamieson, 2023). The expectation that humans will reliably detect, diagnose, and correct the failures of increasingly opaque AI systems is both unrealistic and misaligned with decades of evidence from human–automation interaction learnings (Bainbridge, 1983; Sheridan & Nadler, 2006).

In high-risk domains, this over-reliance on human supervision is particularly problematic. Oversight mechanisms such as "human-in-the-loop" or "human-on-the-loop" often assume that the human will have: a) sufficient time to intervene; b) adequate information to understand what the system is doing; c) the authority and capability to override the system; and d) the cognitive resources to manage the additional workload imposed by supervising complex automation. In practice, these assumptions often break down when systems operate at speeds or levels of abstraction that exceed human cognitive capacity, or when the human-machine interface fails to convey critical information at the right time. Furthermore, supervisory control is not simply a matter of design. It must coexist with real operational constraints such as attention fragmentation, competing goals, and interruptions in dynamic work environments.

Taking humans as the main safeguard against AI failure is thus creating new vulnerabilities compared to current approaches in safety which emphasize the diverse nature of socio-technical systems (e.g. Abbas & Michael, 2025) and overrides the required redundancy measures.

2. Introduction

Together, the challenges with risk definitions, the safety and compliance relations, and human-AI interaction, indicate that efficient safety approaches within high-risk industries demand multi-disciplinary and multi-layered approaches such as barrier management (e.g. Hauge & Øien, 2016). We need to combine regulatory requirements, subject matter expertise, contextual knowledge, human behaviour information, technical knowledge, and technology qualifications techniques to achieve resilience, sustainable and safe operations.

There are several gaps between traditional safety risk frameworks and the risk interpretation provided in the EU AI Act. A summary is presented in table 1.

Table 1. Gaps between safety risk and EU AI Act risk interpretations

Topics	Safety risk	EU AI Act
Scope	Systemic risk	Individual risk
Focus	Context of application	AI systems
Human interaction	Integrated socio-technical systems approach	Isolated human-AI interaction
Risk mitigation	Barrier management approach	Over-reliance on human oversight

Compliance with the EU AI Act does not guarantee that an AI system meets the stringent requirements for technology qualification and barrier management in safety-critical contexts. This has considerable implications including the underestimation of AI-related risks; the isolated assessment of AI systems and models; and limited consideration of human factors aspects. To address these challenges, we explore how performance based and context-aware evaluation frameworks could complement compliance-based assessments. We discuss that in safety critical domains, the existing frameworks for safety risk assessment need to be broadened to include new safety risks and challenges specific to AI and emerging technologies. It will not be enough to rely on "add-on" and compliance-based approaches aligning with requirements provided by the EU AI Act. There is a need to integrate new risk management strategies designed for AI systems with pre-established standards and methodologies for technology qualification, human-machine interaction, and barrier management approaches (Lootz, et al., 2025).

3. Case-study examples

Different types of AI systems will imply different impacts and mitigation strategies in deployment. Different applications of AI systems (information retrieval, real-time operator support, personalized training systems, information selection, recommendations, etc.) raise different safety and assurance questions. Predictive AI systems (such as anomaly detection of predictive maintenance) are most frequently rule-based systems, have clearer processes for quality assessment, and are more transparent to the users. The same is true for prescriptive AI such as optimisation and decision support algorithms. More recent types of AI, such as generative AI (e.g. large language models, retrieval-augmented generation tools), and even agentic AI (where the AI is autonomous and can perform adaptive planning and even execution of tasks) will have more complex and even unforeseen interactions in the operational context and with the humans.

3.1. *Example with generative AI applications*

Generative AI applications currently permeate the AI landscape in industrial context. Several software packages offer "chatbot" style assistants that can be coupled to the companies own data for enabling specific support. They are to assist with complex tasks such as interpreting operational data, generating procedural guidance, and supporting decision-making.

Using techniques as retrieval augment generation (RAG), large language models (LLMs) can provide rapid access to information and generate context-specific recommendations for operators. Desirable impact of their use includes improved efficiency in information retrieval, decision

support, and enhanced situational awareness through summarisation of complex data.

Under the EU AI Act, such systems are typically classified as low risk, as they do not directly control physical processes and are fully controlled/supervised by a human operator. However, when integrated into workflows that influence safety-critical decisions, their impact can be significant.

Our internal testing procedures for RAG solutions has included multiple methods, such as the use of automated evaluation metrics (Escobar et al., 2026), user testing, and early human factors verification, and task analysis (Ardnt et al., 2026).

We performed hands-on testing of three RAG tools (both internally developed and made available by external partners). The first case is described in detail in (Ardnt et al., 2026) and reports an AI assistant for the creation of detailed operational procedures in drilling and well. The other two tools focused on the support for the retrieval and summarization of data from safety incidents databases. All three tools presented a typical chatbot interaction interface, where the user could freely write prompts to the systems through written text.

We found a clear pattern in the results, with very high error rates on outputs and different types of issues, including the presentation of wrong information, both by contradicting information available in the source documents/databases, and by generating information that did not exist (such as creating new cases, not present in the databases, that would confirm or allow a reply to the user's prompt). The overall assessment results are presented in table 2.

Table 2. Evaluation outcome genAI case

Aspect	EU AI Act classification	Safety-critical assessment
Risk Level	Low-risk	High-risk (due to indirect influence on safety decisions)
Required measures	Transparency notice – inform user they are interacting with AI	Technology qualification, barrier analysis, human factors validation
Testing focus	Generic AI system performance (technical)	Context-aware testing under abnormal and failure scenarios

Our findings highlight the risk of hallucinations or inaccurate outputs that can lead to incorrect operator actions. Increased cognitive workload is also a risk both in situations where outputs are ambiguous and as a general consequence of the required extensive verification that the user has to perform to ensure the information is valid, relevant, and accurate. Trust calibration issues such as over-reliance on AI recommendations over time and less thorough verification of the system's outputs, could bypass safety barriers.

3.2. *Example with machine learning*

Equinor's machine-learning (ML) based platform for predictive maintenance and large-scale condition monitoring continuously analyses sensor and process data to detect anomalies early, helping reduce failures, downtime, and operational risk (Equinor, 2026).

Today, the platform includes around 700 deployed ML models and is used on all Equinor offshore installations, monitoring data from about 15,000 sensors across 30 offshore installations. It has already helped identify and report over 200 equipment failures at an early stage.

The machine-operator task allocation is a deliberate attempt to take into consideration major accident risks:

- The model detects anomalies (deviant patterns)
- Human experts evaluate every triggered event
- False positives are fed back into the system manually

- Real positives prompt controlled and human-approved interventions

Thus, no diagnosis or actions are performed by the application. This is left to the operators and equipment experts as before. The main change is that the operators are better informed and alerted earlier of any indication of upcoming equipment failure.

From a major accident perspective, a change in the task allocation (e.g. leaving the diagnosis and/or corrective action to the ML application) would require a consequence assessment for existing safety barriers and would also require a detailed human factors mapping to ensure operator oversight and intervention capabilities. This would be critical to avoid equipment shutdowns and deterioration of existing safety barriers. As such, understanding the nature of the tasks and the work design is crucial to determine whether a solution in low or high risks. In this case, the assessment through the EU AI Act framework and the traditional safety approaches has a similar result, but there still was a need to further analyse the context of operation from a safety perspective. Within the EU AI Act, the assessment often can be determined by the type of AI system that is used and its core task.

Table 3. Evaluation outcome ML case

Aspect	EU AI Act classification	Safety-critical assessment
Risk Level	Low-risk	Low risk
Required measures	Transparency notice	Analysis of the tool in context
Testing focus	Generic AI system performance (technical)	Context-aware testing focusing on tasks

4. Discussion

Returning to the main argument in this paper that EU AI Act compliance will not be enough for safety critical domains, we can clearly see the possibility of obtaining a “low risk” rating within the EU AI Act (that would require only that the user is informed that they are interacting with an AI system) can correspond to an elevated risk in a standard quick safety critical assessment due to indirect influence on operational decisions, especially during time-critical scenarios such as start-up, shutdown, or emergency response. Important aspects to consider when assessing these tools:

- EU AI Act assessment does not cover of substitute other safety and technology qualification assessments
- Context-aware testing is essential. The AI system performance must be assessed under realistic operational conditions, including abnormal/failure scenarios.
- Human factors assessments break down the intricacies of human-AI interaction both in initial use and long-term use of the technology. Analyse operator interaction, trust, and error recovery strategies.
- Integration with pre-existing governing processes is necessary. Ensure outputs are treated as advisory, with clear protocols for verification and override.

As we have seen, the EU AI Act should be taken as a new regulation to follow for AI systems in safety critical contexts – this comes *in addition* to pre-existing regulations on technology qualification, barrier management, and major accident prevention, as well as to the expected upcoming regulations and standards in the near future, covering AI system specific challenges.

Main topics in safety risk approaches are absent of the EU AI Act and as such, developer teams, technology developers, and deployers should not be mistaken by “low risk” labels within this legislation. It does not excuse further testing and assessment for safety risks in safety critical domains such as oil and gas.

For industries with major accident potential compliance with the EU AI act is therefore not sufficient. The Act provides a horizontal, society-level risk framework, useful for governance, compliance, and rights protection, but it does not account for barrier degradation, drift in models and human–machine coordination,

emergent systemic failure modes, nor high-hazard operations where consequences are catastrophic.

AI systems influence detection, diagnosis, decision-making, coordination, and human performance. This means AI can add new layers of complexity that challenge existing barriers. So even if many AI systems in oil and gas fall into "Minimal" or "Limited" EU risk categories, their major-accident relevance may still be high.

An argument might emerge that treating all safety-critical AI systems as high-risk under the EU AI Act could be enough to address these challenges. Our view is that it will not entirely cover it. While it would strengthen governance, it would not replace the need for functional safety assessments, technology qualification, and context-specific testing. Moreover, the consequences of such an approach include increased cost and complexity, potential delays in technology deployment, and would not be bearable nor feasible for most (if not all) technology deployers. A more practical solution is to adopt a hybrid framework that combines EU AI Act principles with established safety standards and regulations at the international, national and even at each company's governing processes. As we highlighted, the context of application of the AI systems (where, for what, and how it is integrated into pre-existing processes) will be essential to determine its safety and adequate mitigation measures.

4.1. *Recommendations*

Key aspects going forward include technology qualification processes as well as human factors validation and verification, where context is seen as crucial and multidisciplinary approaches are required. Risk cannot be assessed solely at the level of the AI model or application; it must consider the operational environment, integration with other systems, and human interaction. Our main recommendations are to:

- Adopt a risk-based approach: Compliance with the AI EU Act alone is insufficient to meet domain specific regulations. We must demonstrate and document safe and reliable use of AI solutions tailored to context
- Risk Assessment must address context of use: Limiting the risk assessment to the AI models or AI applications will not be sufficient to assure safety. AI solutions to be used in safety critical contexts must be developed, tested and assessed with reference to actual use, addressing technical, operational and organizational risks (Havtil Framework Regulations § 11)
- Integrate safety assurance protocols into existing processes: Develop assurance protocols for the development, risk assessment, testing, deployment, use and maintenance of AI solutions. These should be integrated into current workflows and governing documents.
- Adopt a screening approach: Identify applications for safety-critical contexts that require thorough risk assessment and testing to ensure effective risk management.
- Prioritize testing: Conduct tests that evaluate AI performance and reliability in the specific context of use. User testing is crucial to ensure that people can effectively respond to AI errors or failures.
- Provide clear guidance for use: For generic AI applications, include a clear description of intended and foreseeable unintended uses, and provide proper user guidance. This ensures AI systems are utilized appropriately according to established risk levels.

In summary, the EU AI Act introduces important governance principles, but it cannot be viewed as a standalone solution for managing AI risks in safety-critical contexts. Compliance should be treated as a starting point, not the final measure.

6. Conclusion and Further Work

The introduction of the EU AI Act represents a significant step towards structured AI governance. However, this work demonstrates that compliance alone is insufficient for ensuring safety in safety critical domains. The Act's risk

classification framework does not fully capture the complexity of socio-technical systems where AI interacts with safety barriers, operational processes, and human decision-making. As a result, AI systems deemed low-risk under the EU AI Act may still pose substantial hazards when deployed in safety-critical contexts. Future steps in this work will include the drafting of frameworks that integrate new challenges and risks linked to AI systems into pre-existing safety and risk management frameworks in safety critical domains. A multi-disciplinary approach that is context-aware will be crucial to enable efficient use of AI systems in safety critical domains such as oil and gas.

References

Abbas, R.& Michael, K. (2025) Socio-Technical Theory: A review. In S. Papagiannidis (Ed), TheoryHub Book. Available at https://open.ncl.ac.uk / ISBN: 9781739604400

Aderamo, A. T., Olisakwe, H. C., Adebayo, Y. A., & Esiri, A. E. (2024). *Behavioral safety programs in high-risk industries: A conceptual approach to incident reduction*. *Comprehensive Research and Reviews in Engineering and Technology, 2*(1), 64–82. https://doi.org/10.57219/crret.2024.2.1.0062

Ardnt E., Fannemel, R., & Fernandes, A. (2026). Safety insights on generative AI assistants: a case study in drilling and well intervention. EAGE – Digitalization, Stavanger 9-12 March.

Cocklin, T. (2012). Pre-accident investigations. CRC Press, USA.

Endsley MR. Ironies of artificial intelligence. Ergonomics. 2023 Nov;66(11):1656-1668. doi: 10.1080/00140139.2023.2243404. Epub 2023 Aug 6. PMID: 37534468.

Equinor (2026) Machine learning against the machine. https://www.equinor.com/energy/machine-learning.

Escobar et al. (2026). Evaluating Retrieval-Augmented Generation for Subsurface Applications: Lessons Learned. EAGE – Digitalization. Stavanger 9-12 March.

European Commission. (2024). Artificial Intelligence Act. Brussels: EU Publications.

Hauge, S. & Øien, K. (2016). Guidance for barrier management in the petroleum industry. SINTEF Digital: Research Report. Norway 35ef4d1e-d8fe-4b1b-bfef-af9d6fa9cbac

IEC. (2010). IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems. Geneva: International Electrotechnical Commission.

ISO. (2018). ISO 31000: Risk Management – Guidelines. Geneva: International Organization for Standardization.

Kyriakou, K., Otterbacher, J. In humans, we trust. *Discov Artif Intell* **3**, 44 (2023). https://doi.org/10.1007/s44163-023-00092-2

Lootz, E., Bergh, L.I.V., & Heide, B. (2025). Risk management and uncertainty of artificial intelligence in a high hazard industry. In *Proceedings of the 35th European Safety and Reliability& the 33rd Society for Risk Analysis Europe Conference*. 15-19 June, Stavanger, Norway. doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P2957-cd

Markusen, C., van der Meulen, M., van de Merwe, K., & Kvinnesland, K. (2024). State of the art for responsible use of artificial intelligence in the petroleum sector. DNV Soler Garrido, J., De Nigris, S., Bassani, E., Sanchez, I., Evas, T., Andre, A.A., & Boulange, T. (2024). Harmonised Standards for the European AI Act. JCR139430. European Commission. JRC Publications Repository - Harmonised Standards for the European AI Act

NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Norsk Industri (2025). Safety, leadership, and Learning: A practical guide to HOP, 2nd Edition. norsk-industri-hop-veileder-2026-engelsk.pdf

NORSOK. (2024). Z-013: Risk and Emergency Preparedness Analysis. Oslo: Standards Norway.

OECD (2024). Defining AI incidents and related terms. Organization for Economic Cooperation and Development Publishing. Defining AI incidents and related terms (EN)

Ringstad, A.J., Fernandes, A., & Ludvigsen, J.T. (2025). Artificial Intelligence and Major Accident Risk: Towards a Framework for Systemic Safety Risk Assessment. In *Proceedings of the 35th European Safety and Reliability& the 33rd Society for Risk Analysis Europe Conference*. 15-19 June, Stavanger, Norway. doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P2957-cd

Sheridan, T., & Nadler, E. (2006). *A review of human-automation interaction and lessons learned* (DOT-VNTSC-NASA-06-01). United States. National Aeronautics and Space Administration; John A. Volpe National Transportation Systems Center. https://rosap.ntl.bts.gov/view/dot/8879

Skraaning, Jr., G. & Jamieson, G.A. (2023). The failure to grasp automation failure. Journal of Cognitive Engineering and Decision Making, vol.18 (4). https://doi.org/10.1177/15553434231189375