AI TECHNOLOGY

“State of the art” for responsible use of artificial intelligence in the petroleum sector

Norwegian Ocean Industry Authority (Havtil)

Report no.: 2024-1519, Rev. 1
Document no.:
Date: 20.12.2024

Project name:	AI technology	DNV Energy Systems
Report title:	"STATE OF THE ART" FOR RESPONSIBLE USE OF ARTIFICIAL INTELLIGENCE IN THE PETROLEUM SECTOR	Digital Technoloy, Høvik Veritasveien 1, 1322 Høvik Tel: +47 67579900
Customer:	Norwegian Ocean Industry Authority (Havtil) Professor Olav Hanssens vei 10 4021 Stavanger	945748931
Contact person:	Linn Iren Vestly Bergh	
Date:	20.12.2024	
Project no.:	10434783	
Org. unit:	Digital Technology, Subsea, Risers and Pipelines	
Rapport no.:	2024-1519, Rev.1	
Document no.:		

Delivery of this report is subject to the provisions of the relevant contract(s):

Mission description:

Havtil has commissioned DNV to prepare a state-of-the-art report that covers the basic risk factors associated with the development and use of artificial intelligence (AI) in petroleum industry, particularly with regard to major accident risk.

Performed by:	Verified by:	Approved by:
Digitally signed by Christian Markussen Date: 2024.12.20 09:14:30 +01'00' Christian Markussen Global Practice Lead	Agrell, Christian Digitally signed by Agrell, Christian Date: 2024.12.20 10:00:03 +01'00' Christian Agrell Principal Researcher	Jahre-Nilsen, Per Digitally signed by Jahre-Nilsen, Per Date: 2024.12.20 11:41:20 +01'00' Per Jahre-Nilsen Head of Section
Meine van der Meulen Digitally signed by Meine van der Meulen Date: 2024.12.20 12:24:43 +01'00' Meine van der Meulen Senior Principal Researcher		
Digitally signed by Kvinnesland, Kenneth Date: 2024.12.20 09:49:23 +01'00' Kenneth Kvinnesland Senior Principal Consultant		
van de Merwe, Koen Digitally signed by van de Merwe, Koen Date: 2024.12.20 09:24:29 +01'00' Koen van de Merwe Principal Researcher		

Internally at DNV, the information in this document is classified as:

☒ Open	--	--
☐ DNV Restricted		
☐ DNV Confidential		
☐ DNV Secret		

Keywords: Artificial intelligence, major accidents, petroleum sector

Rev. no.	Date	Reason for issue	Prepared by	Verified by	Approved by
0	2024-10-25	First edition in Norwegian	Christian Markusen, Meine van der Meulen, Koen van de Merwe, Kenneth Kvinnesland	Christian Agrell	Per Jahre-Nilsen
1	2024-12-16	Version in English	Christian Markussen	Christian Agrell	Per Jahre-Nilsen

Table of contents

1 SUMMARY

The Norwegian Ocean Industry Authority (Havtil) has commissioned DNV to prepare an overview over state-of-the-art and best practice that covers the risk factors associated with the development and use of artificial intelligence (AI) in petroleum industry, particularly regarding applications with potential for a major accident. This summary briefly describes the most important observations in the report.

Where is AI expected to be introduced?

DNV expects AI to be introduced in the petroleum sector at a rapid pace. Initially, AI will be used to generate different types of documents and source code and will in the short term also be used in different types of advisory applications, e.g. within operational optimization and predictive maintenance. In the longer term, AI is also expected to be used in control functions, for example related to lifting operations, and within drilling and well control. Autonomous AI-based systems will initially be introduced where the potential for damage is low, for example in underwater vehicles and drones.

Risks associated with the use of AI

This report identifies several AI-related risks that can cause operations to end up in an unsafe state. Some examples are insufficient training of algorithms, poor data quality, model degradation over time, and overfitting to training data. In addition, AI-based systems will be vulnerable to weaknesses in software design, hardware failure, and malicious actions such as cyber-attacks. As described above, AI is expected to be used in different types of systems and operations and, for most of these, there is a risk that information generated using AI-based systems can lead to incorrect operational decisions, which in turn can lead to an accident. This report therefore uses the term "safety-related systems" as a collective term that includes safety systems, control and monitoring systems, as well as advisory, planning, and condition-monitoring systems.

Use of barriers

It is expected that in the future, the safety of people and the environment will continue to be managed through the barrier philosophy that forms the basis for Havtil's regulations. The thinking behind the barrier philosophy is that no matter how well one tries to achieve a safe and robust solution, failures, hazards, and accident situations will occur, and then barriers must come into effect and control such situations. If an AI-based application produces results that cause an operation to end up in an unsafe state, this philosophy dictates the use of barriers to prevent the condition from escalating into a dangerous event. Examples of such barriers are human override, the use of independent control functions, the use of independent safety functions that shut down the process being controlled, and the use of operational restrictions that reduce the risk of a dangerous incident if the other barriers are not effective.

The interaction between AI and people

For many types of operations, it will currently be humans who make the decision to activate barriers of the type described above. Part of the challenge is the fact that unwanted conditions in software do not always lead to alarms. For example, it is not a given that an alarm will be triggered if an AI algorithm is exposed to an operational scenario that it is not trained for. In such a situation, barriers will only be activated if people are able to understand that something is wrong, based on the available information. For this reason, this report also focuses on the interaction between AI and people, the need for human-centred design of the AI solution, and the need to be able to detect unsafe conditions in a way that is independent of the system that contains AI.

Independent human detection of unsafe conditions may in some cases be possible by comparing results produced by AI with results produced using an alternative technology. Measurements of process parameters can also be used to identify unsafe conditions caused by AI but cannot be used alone to detect all forms of unsafe conditions. This is because common cause failures in software or problems with data quality can negatively affect both the decision-making processes that take place in software and those of the operator. Detection based on human observation of process parameters and other related information, requires there to be enough time to to make good judgments, and

deep insight into the process being controlled. Such insight will in most cases be more important than insight into how AI works.

Regardless of the level of knowledge, it may be difficult to identify abnormal situations manually if the deviations from the normal process are relatively small, which may be a problem if one also operates with small margins against unsafe process conditions.

AI will typically increase both the capability and complexity of a system. Research shows that the more capable and complex a system becomes, the less able the user is to understand the system and its limitations and to monitor it reliably. The challenges associated with human detection of unsafe conditions indicate that the industry should explore the possibilities for automatically detecting potential unsafe conditions caused by AI.

Factors that may limit where AI can be introduced

For a barrier to be effective, it must be possible to detect that an unsafe condition has arisen regardless of what has caused it - otherwise it cannot be trusted that the barriers will be activated when needed. If software has somehow caused an unsafe state, detection for whatever reason is only possible if there is access to a detection mechanism that is completely independent of that software. A safety function that is automatically activated when its measurements shows that one of the safety critical process parameters has exceeded a predefined threshold, is an example of a barrier where such independent detection is available.

However, increased levels of automation may in many cases make it difficult to detect an unwanted condition in an independent way. This means that the industry is moving towards a grey area where there, like in some other industries, are critical and complex functions that must continuously work as intended, to ensure safety. These challenges will often be present regardless of whether AI is being used or not, but the use of AI can exacerbate the problem.

Due to the barrier philosophy where separation of control and safety functions is required, the requirements for how to develop, verify, and validate software-based control functions are not particularly strict. This means that the threshold for using AI in control functions may be lower than the threshold for using AI in safety functions. Lack of mechanisms that can detect unsafe conditions without being dependent on the system using AI, is nevertheless expected to limit where AI can be introduced in a safe way.

The standards that Havtil's regulations refer to for safety functions set very strict requirements for how software is to be developed, verified, and validated, and the introduction of AI in such functions will thus trigger a considerable burden of proof. For this reason, AI is not expected to be introduced in such functions in the near future. However, it cannot be ruled out that AI-based components may be introduced in the longer term - for example, one can imagine that AI-based activation of safety functions comes as an addition, where today there is only human activation.

For some AI systems, the results produced may not be deterministic. This means that if several tests are performed with the same input data, the results are not necessarily the same, which can make it difficult to qualify, validate, and maintain software that contains AI. This can also limit where AI can be introduced, and the industry should explore how this challenge can be solved.

Need for common guidelines for the use of AI in the petroleum industry

Havtil's regulations refer to a wide range of standards and guidelines, both Norwegian and international. Most companies choose to follow these, since otherwise they must demonstrate that alternative approaches are just as good or better. The use of common standards and guidelines contributes to a harmonized level of safety in the petroleum industry. However, so far, no standards or guidelines have been prepared specifically for the safe use of AI in the petroleum industry. To maintain a harmonized level of safety and reduce the burden of proof on the individual companies, it would be beneficial if relevant players in the industry could join forces to prepare guidelines that represent best practice for the use of different types of solutions containing AI.

Such an effort should also make it easier to meet the requirements of the EU AI Act. Currently, individual players that want to use AI must specify and respond to the requirements of the regulation in their own management systems. Such operationalization of high-level requirements is usually labor-intensive, but it should be possible to reduce the workload on each individual organization if the industry makes a joint effort.

2 INTRODUCTION

2.1 Project description

Havtil has commissioned DNV to prepare a knowledge overview that covers the basic risk factors associated with the development and use of artificial intelligence (AI) in the petroleum industry, particularly regarding the risk of major accidents. The purpose of the study is to increase knowledge about the risks associated with the development and use of AI in operations that are important for safety on the Norwegian continental shelf.

The knowledge overview shall contribute to increased understanding of the risks associated with AI systems in the petroleum sector. It shall explore how AI can improve both efficiency and safety, while considering the unique risks AI introduces compared to traditional IT and automation systems.

It is necessary to use up-to-date literature and knowledge about AI models and their potential consequences for safety in the petroleum industry. This includes issues such as data quality, integrity, training, re-training, testing, transparency, and explainability, as well as implementation. It also includes methods and techniques for identifying and managing risk in different phases of the AI model's life cycle, and an emphasis on good situation awareness and the communication of uncertainty that facilitates human-machine interaction. Furthermore, the overview shall highlight relevant international standards, such as those published by ISO, DNV, and NIST, and their application to ensure that AI solutions comply with regulations and achieve a high level of safety.

2.2 Background

The petroleum industry in Norway operates fixed and floating offshore installations, onshore facilities, and an extensive pipeline system. There is considerable potential for various forms of major accidents, for example related to fires, explosions, and oil leaks. For this reason, there is a strong focus on the safety of both people and the environment.

The overall framework for safety work is set out in Havtil's five sets of regulations for oil and gas activities. For each set of regulations, Havtil provides interpretations and guidelines that describe how the regulations can be applied, including references to industry standards that can be used to meet the requirements. It is the players' responsibility to follow the regulations, which are largely formulated as functional requirements with a focus on needs and desired results, rather than prescription of specific solutions. This means the players in the industry in principle have quite a lot of freedom regarding how the requirements are to be met, but the fact that the industry has agreed on the use of specific standards and guidelines that are referred to in Havtil's regulations has nevertheless contributed to a high degree of standardization among the players.

A fundamental principle underpinning many of the requirements in the regulations is that the failure of a component or system or a single mistake shall not lead to unacceptable consequences. There must be independent barriers in place to prevent such incidents from escalating and leading to accidents. Regarding this, Havtil has published a memo related to barrier management /7/ and a note related to risk management /8/ that provide additional guidance in beyond the guidelines contained in the regulations themselves.

AI is now being introduced into society at a relatively rapid pace, posing challenges for the work related to safety in the petroleum industry. So far, no standards or guidelines have been developed specifically for the safe use of AI in the petroleum industry. Existing standards and guidance consider that different types of software may deliver results that are incorrect or not fit for purpose but have not been developed with AI in mind. Thus, there is a possibility that the way software-related risk is currently being handled will not be effective for AI-based functions.

In addition, Norway has become the EU's most important supplier of gas, which means that the ability to continuously deliver gas has become even more important than before. A challenge here lies in the fact that Havtil's regulations focus on safety and, only to a limited extent, on the systems available for production. There are no strict requirements for systems that are only considered critical for the availability of production and deliveries, and the threshold for using AI for systems that are only considered availability-critical will thus be significantly lower than for safety-critical systems.

This can potentially have positive effects in terms of both availability and volume of deliveries, but the risk of AI having a negative impact must be managed. Related to this, petroleum industry players operating infrastructure that is critical for gas supply to the EU must prepare to meet the new regulatory requirements of the EU Regulation on AI (EU AI Act) /50/.

In addition, there has for many years been a trend towards an increasing degree of automation in the industry and decisions being made to a greater extent based on reporting from software that is analyzing large amounts of data. This means the traditional distinction between operational technology (OT) and information technology (IT) is becoming increasingly blurred, and that systems which were previously only considered advisory have become more critical. The introduction of AI is expected to reinforce this trend by making it more difficult to detect erroneous or inaccurate results from software.

The introduction of AI will also challenge the methodology for technology qualification since results from AI-based software may not be deterministic. This means that if you carry out the same test several times, you may get different results, which poses challenges related to verification and validation.

In sum, this means that the petroleum industry has a great need to develop guidelines for how AI-related risk should be managed for different types of activities, and this knowledge overview is intended to contribute to this work.

2.3 Method

An extensive literature search has been carried out in reputable databases and peer-reviewed journals to find relevant articles, reports, and books on AI and safety aspects in the industry. Keywords such as "artificial intelligence", "major accident risk", and "safety-critical systems" were used to ensure adequate results.

Furthermore, we have studied national and international regulations as well as standards and guidelines that are relevant to the use of AI in the petroleum industry. An important part of the method was the identification and evaluation of relevant standards and frameworks developed specifically for AI, but we have also identified relevant standards and guidelines that were not developed with AI in mind.

We also assessed existing practices for risk management and the use of barriers. This included assessments of how current strategies for managing software-related risk can be adapted for the implementation of new AI-based solutions.

AI in the form of a language model has also been used in the work, but everything in the final version has been reviewed and quality assured by humans.

2.4 Scope and limitations

Since this study is about the use of AI in the petroleum sector as an industry with the potential for major accidents, this report focuses on the safety of humans and the environment in the same way as Havtil's regulations.

This is an important delimitation, since it means that the focus of the report is barrier philosophy when using AI. A report where the focus is on achieving the highest possible operational reliability and value creation using AI solutions, would have a different structure.

2.5 Definitions

English term	Definition
Barrier element	A barrier element means a technical, operational, or organizational measure or solution that is part of the realization of a barrier function. Technical barrier elements are equipment and systems that are part of the realization of a barrier function. Organizational barrier elements are personnel with defined roles or functions and specific expertise that are part of the realization of a barrier function. Operational barrier elements are the actions or activities that personnel must perform in order to realize a barrier function.

English term	Definition
Explainability	Ability of an AI system to express important factors that affect the AI system's performance in a way that humans can understand (ISO/IEC 22989, § 3.5.7).
Generative AI	Refers to models used to generate new data. These models aim to learn the underlying distribution of the data so as to produce new examples similar to the training data.
Information technology	The use of computers and telecommunications equipment to store, retrieve, transmit, and manipulate data. This incorporates the handling and processing of data, including hardware, software, networks, and databases that enable information processing.
Artificial intelligence	A system's ability to perform functions generally related to human intelligence such as reasoning and learning ability (ISO/IEC 2382-28).
Life cycle	The evolution of a system, product, service, project, or other man-made entity, from its creation to its decommissioning (ISO/IEC 22989, § 3.1.22).
Operation technology	Hardware and software used to detect or cause changes through the direct monitoring and control of industrial facilities.
Robustness	The ability of a system to maintain its level of performance under all circumstances (ISO/IEC 22989, § 3.5.12).
Bias	Systematic discrepancies or unfairness in the results of an AI model, often as a result of the data used for the training or design of the model.
Transparency	Functionality that provides users with appropriate information about the AI system's situation awareness and planned actions.

2.6 Abbreviations

Abbreviation	Definition
HCD	Human-centred design
IEC	International Electrotechnical Commission
ICT	Information and Communication Technology
ISO	International Organization for Standardization
IT	Information Technology
KI	Artificial Intelligence
LLM	Large language models
ML	Machine Learning
OT	Operation technology
RRF	Risk Reduction Factor
V&V	Validation and Verification

3 INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) encompasses a range of technologies designed to mimic human cognitive functions, with the ability to learn, reason, solve problems, and make decisions. For the Norwegian offshore industry, AI applications may help to reduce costs and contribute to advances in operational safety, efficiency, and decision support.

3.1 What is AI?

There are many definitions of AI. Table 3-1 lists some definitions taken from relevant standards that show there is quite a widespread in how it is defined.

Table 3-1 Definitions of AI.

Standard	Definition
ISO/IEC 2382:2015, ISO/IEC 29140:2021, ISO/TR 24291:2021	Branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement.
NS 11041:2024, ISO/IEC 2382-28:1995	A system's ability to perform functions that are generally related to human intelligence such as reasoning and learning ability.
DNV-RP-0671	Computer programs that generate output based on learnt transformations of data or other forms of automated reasoning.
National strategy for artificial intelligence	Artificially intelligent systems perform actions, physical or digital, based on the interpretation and processing of structured or unstructured data, with the aim of achieving a given goal.
EU AI Act, EU Artificial Intelligence Regulation	AI system means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

These different definitions of AI have some common elements, but also address different aspects of AI. The essence and commonalities of these definitions are described below:

- **Human intelligence as a reference**: several definitions refer to functions that are associated with human intelligence, such as reasoning, learning, and problem-solving. This is a recurring theme indicating that AI is about mimicking human cognitive abilities.
- **Autonomy and adaptation**: several of the definitions (notably those in ISO/TR 24291 and the EU AI Act) highlight the ability of AI systems to operate autonomously and adapt to new situations based on learning from data. This means that AI systems not only follow predefined rules but can also adapt based on experience.
- **Data management and learning**: AI systems are related to the processing of data, whether structured or unstructured, and the use of this data to generate insights, recommendations, or actions. Learning/training from data is central to several of the definitions.
- **Purposefulness**: the definitions also highlight that AI has a purpose or goal that the system is trying to achieve, whether this is explicitly specified or implicit.

Several of the definitions, especially those in the National Strategy for Artificial Intelligence and EU AI Act, are broad enough to also include digital systems that would not traditionally be seen as AI, such as simulation models. This is due to the focus on the ability of systems to interpret data and influence their environment, which can include a number of other digital technologies.

In principle, this knowledge overview considers everything that falls under one of the definitions above but focuses on the types identified in chapter 3.2 below, since these are considered most relevant for systems with major accident potential.

3.2 Types of Artificial Intelligence

Some examples of types of AI relevant to this knowledge overview:

- **Machine learning (ML)** enables systems to learn from data, identify patterns, and make decisions with minimal human intervention. Relevant examples are ML applications for predictive maintenance, where algorithms predict specific types of equipment failures before they occur, and ML applications for anomaly detection in large amounts of operational data,
- **Generative AI** refers to models used to generate new data. These models aim to learn the underlying distribution of the data to produce new examples similar to the training data.
- **Discriminative AI models** are a type of artificial intelligence used for prediction and inference based on existing data. They do not generate new data but focus on identifying patterns and making decisions based on those patterns.
- **Large language models (LLMs)** are a type of generative AI that is trained on vast amounts of text data to be able to understand, generate, and process natural language in a human way. Models such as GPT (Generative Pretrained Transformer) use deep neural networks to capture patterns, grammatical rules, meanings, and context in language.
- **Rule-based models** include systems that follow predefined rules or logic to make decisions, potentially in combination with machine learning. For example, rule-based systems can monitor operational parameters in real time and send out alerts or take corrective action when anomalies are detected.
- **Hybrid AI models** integrate data-driven approaches with basic physical principles (deterministic or statistical) to model and predict system behaviour. Examples include the optimization of drilling operations based on geological models and real-time data, and the dynamic simulation of how offshore structures will respond to impacts from the operational environment.

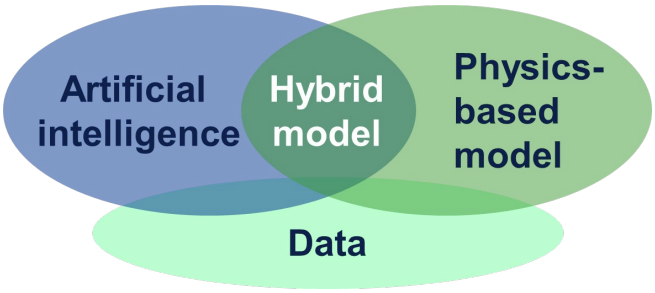

Figure 3-1 Relationship between AI and physics-based models.

Note that the applications mentioned above focus on AI used in direct support of operations. There are a few other applications that will be relevant to the petroleum industry. For example, it must be expected that AI will be used to automatically generate source code for use in both control and advisory systems.

The different types of AI offer toolkits that have the potential to improve safety and operational efficiency in the petroleum industry. Advances in ML and LLMs open new opportunities in data analysis and decision support, while rule-based and physics-based AI applications support decision-making based on established knowledge and principles.

3.3 AI in Norway

AI is expected to play an increasingly important role in Norway, and in 2020 the government prepared a national AI strategy /49/ which aims to promote innovation and the responsible, safe use of AI systems. The strategy focuses on transparent use, data security, and the ethical development of AI technologies, among other things.

The Norwegian government has allocated NOK 1 billion to fund research on artificial intelligence and digital technology. This money will contribute to greater insight into the consequences of technological development for society. It will also provide more knowledge about new digital technologies and opportunities for innovation in business and the public sector. The initiative is financed within the framework of the Ministry of Education and Research budget. The research funding will have three main tracks:

- society-related research, such as legal regulation, demography, privacy, etc.
- research on digital technologies, such as next-generation ICT, digital security, data quality, etc.
- research on how digital technologies can be applied to innovation in private and public industry.

There is already several established centre for AI research, and one billion NOK is planned to fund more centres. Examples of already established centres are NorwAI, NAIL, CAIR, dScience, BigInsight, NORA, and Integreat.

4 RISKS AND VULNERABILITIES RELATED TO THE DEVELOPMENT AND USE OF AI

This chapter illustrates that many types of systems may have an impact on safety, even if they are not defined as dedicated safety systems. This is important since the strict requirements set for dedicated safety systems mean that the threshold for introducing AI in such systems will be very high, while it will be significantly lower in other types of safety-related systems. Furthermore, this chapter looks at the use of AI in a systems perspective, where the system consists of the people, technology, and organization. It identifies typical applications where AI is expected to be used in the relatively short term, and identifies risks associated with these.

4.1 Safety-related systems

In this document, the term safety-related system is used, as illustrated in Figure 4-1. This refers to programmable logic systems and not to mechanical systems such as a pressure protection valve. Although the safety-related systems are primarily responsible for the safety functions, the other systems could also affect the potential for a major accident if they do not work as intended. The figure distinguishes between time-critical and non-time-critical systems, where time-critical systems refer to processes or tasks that require rapid action or execution within a limited time frame, while non-time-critical systems do not require rapid action. Time-critical systems will typically be implemented as part of the operations technology (OT) system, which includes process control, emergency shutdown, and safety functions.

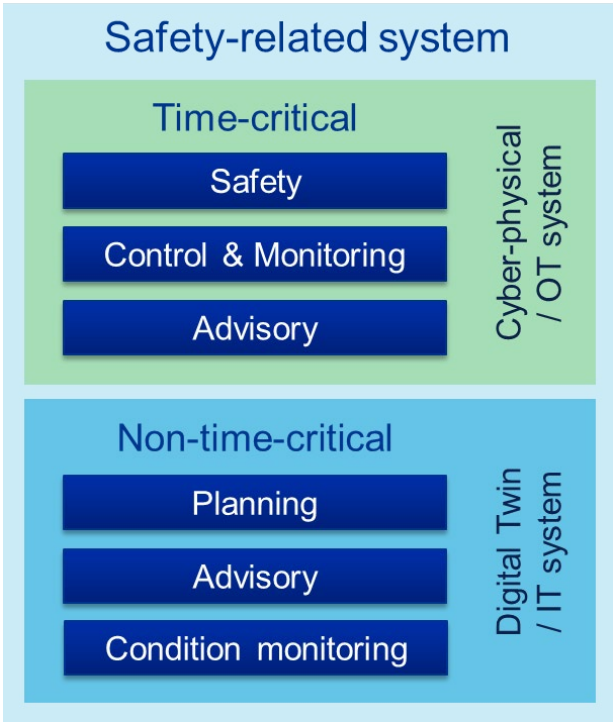

Figure 4-1 Safety-related systems

- **Safety system** – a programmable system that integrates detection mechanisms, logic devices, and actuators with a view to detecting hazardous conditions and bringing a process or system to a safe state. These systems are designed to reduce the risk of dangerous incidents by automatically or manually initiating predefined actions, such as emergency shutdown or depressurization. Relevant standards here include NORSOK S-001, IEC 61508, IEC 61511, Offshore Norway's guideline 070, ISO 13702, and ISO 13849, and these set strict requirements for how such systems are to be designed, verified, and operated.

- **Control and monitoring system** – an automated system that controls, monitors, and regulates processes and operations. Examples are process control systems, drilling systems, and other automation systems. Relevant standards include NORSOK P-002, NORSOK I-001, NORSOK I-002, and ISO 10418.
- **Advisory system** – a system that provides decision support by analysing data to support the operator's situation awareness and suggest further actions. Advisory systems can be used to optimize operations, predict maintenance needs, and improve safety and efficiency. These systems can combine expert knowledge with data analysis to give operators and managers better insight into a situation and recommend actions. Advisory systems can be used in both time-critical and non-time-critical systems, but there are few established standards for this kind of systems.
- **Planning system** – a system used to plan operations and maintenance that have the potential for accidents, such as drilling operations, lifting operations, maintenance at the processing plant, etc. This involves coordinating all activities to avoid unforeseen problems and ensuring that all resources are used efficiently while ensuring safety. There are few established standards for this kind of system.
- **Condition monitoring system** – a system that continuously collects, analyses, and reports data from various processes and equipment. These systems are used to detect anomalies and faults, analyse trends, etc., to predict when equipment will fail to plan maintenance or replacement. This is intended to identify possible problems at an early stage and ensure continuous operation and safety. DNV-RP-A204 and ISO 13374 are examples of a relevant recommended practice and a relevant standard.

4.2 AI in a systems perspective

In an industrial perspective, AI will often be integrated into a system that consists of a mechanical system, a digital system where AI can be included, and people who operate the system. In addition, the organizations that develop and maintain the system, including the governance system, will affect how safe and reliable it is in operation. Within the petroleum sector, this is referred to under the term HTO (Human, Technology, and the Organization):

- **Human**: this focuses on human factors such as competence, experience, the work environment, and how human errors or successes can affect the safety and efficiency of the organization. It is also about how people interact with technology and organizational structures.
- **Technology**: technology refers to the equipment, machinery, software, and systems. The technology perspective looks at how the technology is used, how it affects work processes, how it can affect the situation awareness of the operator, and how it can be optimized for safety and efficiency.
- **Organization**: this refers to organizational structures, procedures, culture, and leadership. Organizational factors can have a major impact on both people's performance and how technology is used. Effective communication, clear responsibilities, and good routines are important aspects here.

The HTO perspective is particularly important for developing holistic systems that are safe and robust. Regarding the technology part, it makes sense to divide this into mechanical and digital systems as these are typically developed in different settings, and it is critical to understand and analyze the interaction between them, see Figure 4-2. The interaction between these elements, where AI is also integrated, makes this a complex system.

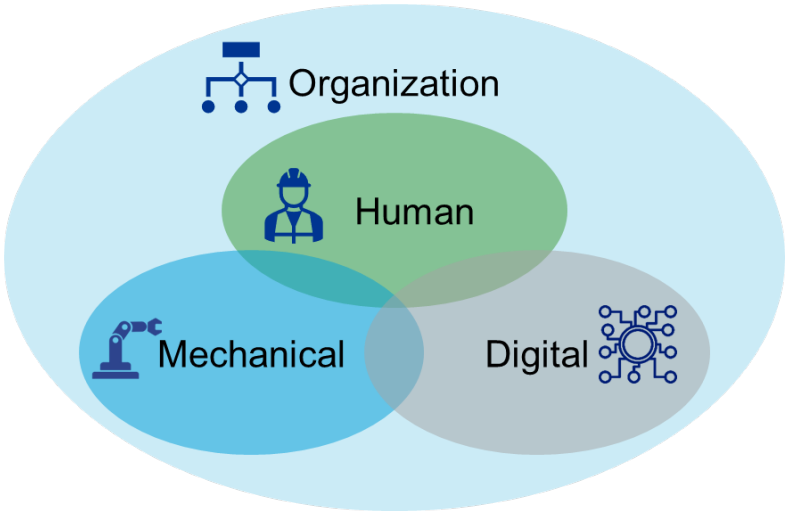

Figure 4-2 The interaction between human, technology, and the organization.

In principle, AI can be part of all safety-related systems (see Figure 4-1) but the measures to qualify AI (the solution) must be proportional to the criticality of the system, and for this reason DNV considers it unlikely that AI in the short term will be introduced in what is classified as safety systems in the petroleum industry.

It will not be sufficient to qualify the AI model as an independent module, it must be qualified based on the requirements for the system as a whole, and in some cases the qualification of AI may also drive changes in the system. This also includes requirements to communicate to the users of the system whether the information produced has sufficient quality and credibility.

4.3 Use of barriers

Barriers are measures that are intended to prevent hazards or accidents from happening, or to reduce the consequences if something does happen. Barrier elements can be technical (such as equipment and digital systems), organizational (management models and procedures) or operational (human actions).

A bowtie risk model is often used to visualize the risk of major events, both physical and digital. It systematically identifies threats and barriers (left side) to an unwanted event (centre), and barriers that can prevent the escalation of the incident and thus reduce the consequences (right side). Figure 4-3 shows a simplified version of such a bowtie model for AI.

Havtil sets requirements for barriers and barrier management, which is about ensuring that all barriers function properly throughout the entire life of an installation or operation. Barrier management (based on safety barriers and the barrier management note) is about:

- identifying threats and impairments
- defining the necessary barriers to deal with the threats
- maintaining, testing, and validating barriers to ensure they function as intended
- following up and monitoring the condition of the barriers, both technically and organizationally.

Performance influencing factors refer to conditions that can affect how well a barrier works, i.e. its ability to perform its intended function under normal or critical conditions. Barrier management is about understanding and controlling these factors so that the barrier's performance is not impaired. Some performance influencing factors for barriers are:

- technical factors, such as wear and tear, maintenance, aging of equipment, sensor failure, etc
- human factors, such as knowledge and skills, fatigue and stress, time available for action, situation awareness, and communication
- organizational factors, such as leadership, culture, procedures, policies, and resources
- environmental factors, such as environmental loads, a corrosive environment, temperature, etc., that can affect performance
- interaction between barriers - for example, a failure in a technical barrier may require an operational barrier, such as a very quick operator reaction.

AI in industrial systems can both be a potential threat and/or act as a barrier against a threat. In any case, it is essential to evaluate the AI application based on how it affects the system and how it fits into a barrier perspective.

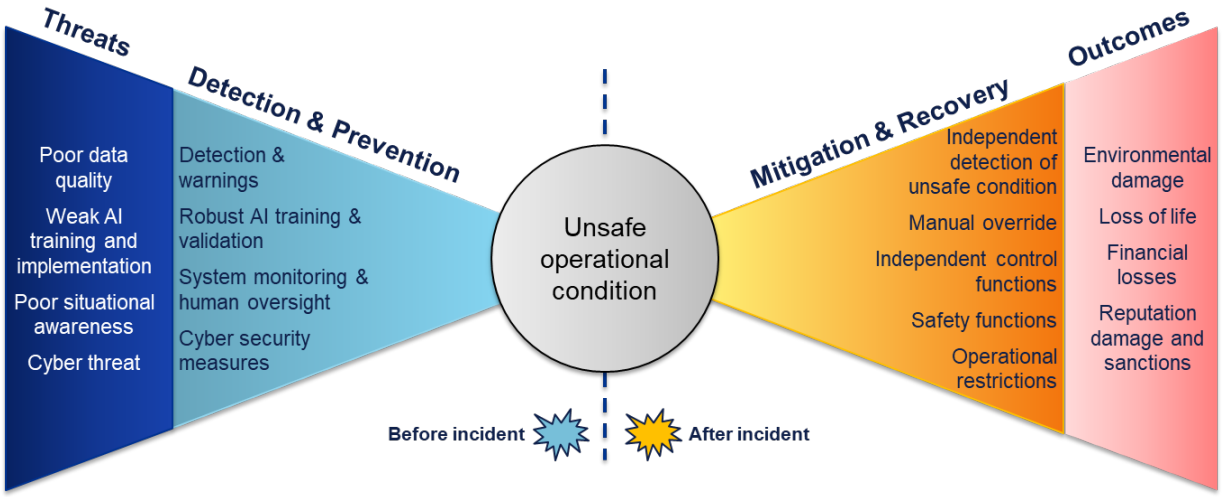

Figure 4-3 Bowtie model for AI systems.

The content of the different elements of the figure is as follows:

Threats (causes that can lead to an undesirable operational state)

Examples of causes that can lead to an undesirable operational state are as follows:

- Inadequate AI training or bias that leads to AI misinterpreting situations or producing uncertain decisions.
- Poor data quality that leads to the AI system making wrong decisions.
- Model degradation - which occurs when the conditions under which an AI algorithm was trained gradually change over time, or when unexpected variables are introduced into the operational environment.
- Overfitting to training data, where an AI model adapts too well to training data that includes noise and details that are not relevant. This means the model performs well based on training data, but poorly with new data.
- Errors introduced in the interaction between AI and humans.
- Errors in the implementation of the application containing the AI algorithm and/or errors in the implementation of other software in the relevant technology stack, e.g. in the operating system.

- Failure in the hardware on which the software containing the AI algorithm is being executed. This can e.g. be due to so-called random hardware failures or negative influences from the environment, such as a rise in temperature or electromagnetic radiation.
- Malicious actions that affect the AI algorithm, for example in the form of hacking, computer viruses, jamming of communication, or sabotage.

Detection and prevention (barriers to prevent an unwanted condition from occurring)

Examples of preventive barriers are as follows:

- Testing and validation of AI models on a regular basis to ensure they work correctly in real-world situations.
- Real-time monitoring and alerts: continuous monitoring of system performance, with real-time alerts for abnormal events.
- Continuous monitoring of incoming data to check data quality.
- Security protection: firewalls, encryption, and multi-layered security protocols to prevent unauthorized access to the AI system.
- Validation of AI-generated information using other tools that generate similar information in a different way. For example, it may be relevant to use different types of simulation tools, with and without AI.
- Use of AI is limited to being advisory, so that people are involved in critical decision-making.
- Uncertainty quantification, which involves the implementation of methods that quantify uncertainty in AI models. This allows predictions to be assigned a confidence level.

Unsafe operational condition

This is the central element of the model (the circle in the middle of the figure). It represents a situation where the AI-based system has delivered data or a command that makes the operational situation unsafe. For example, the operational situation may represent a rare event that the AI system is not trained for.

Mitigation and recovery (barriers that prevent escalation to a dangerous incident)

Examples of such barriers are as follows:

- Manual override allows human operators to override AI decisions at critical moments.
- Use of control functions that are independent of the application that contains AI and are capable of keeping the controlled process within a safe state. Such functions will be activated based on manual or automated detection of the undesirable condition.
- Use of safety functions that shut down the controlled process. Such functions will be activated based on manual or automated detection of the undesirable condition.
- Operational restrictions that reduce the consequences even if escalation occurs. An example of this is that people are physically excluded from the area where an operation is taking place.
- Other measures that reduce the consequences even if escalation occurs, such as firefighting.

Note that manual override using software running on the same hardware as the AI algorithm will not be an independent barrier, regardless of what caused the unwanted condition.

In addition, if they are to be effective regardless of the cause of the problem, all the first three types of barriers will be dependent on being able to detect the undesirable condition independently of the system containing AI. This must also happen so quickly that the barrier is activated in time to have an effect.

For example, it is not a given that you will receive an alarm from the system that contains AI, if the cause of the undesirable event is that the AI algorithm is not trained for an operational scenario that occurs rarely, or if the cause is an implementation error. If the undesirable condition cannot be detected in other ways, then it may escalate to serious outcome.

This detection problem is in many cases expected to be the main challenge when it comes to the safe use of AI, especially when relying on human detection. The issue is not new and will often be present in connection with the automation of processes, regardless of whether AI is used, but AI can in some cases exacerbate the problem. See also chapters 5.2 and 5.7.

Outcome (potential consequences if the adverse event escalates)

Worst-case scenarios will vary greatly depending on the type of AI application, but one can imagine dangerous well incidents, offshore cranes moving in a dangerous way, unwanted spills, etc.

4.4 Examples of applications considered relevant in this study

The use of AI in the energy sector, especially in the offshore industry, is in its early stages but is expected to expand significantly. Some examples of relevant applications are given below:

- **Applications that provide decision support:** at the advisory end of the spectrum, AI systems are designed to support human decision-making, rather than replace it. These applications mainly focus on providing insights, recommendations, and enhanced analytics that help engineers and operators make informed decisions. Examples include:
    - **predictive maintenance:** using AI to analyse data from sensors and machinery to predict equipment failures before they occur, so that maintenance can be carried out in a timely manner and unplanned downtime is reduced /35/ /36/
    - **planning:** using advanced AI analysis to predict maintenance needs, optimize resource allocation, and plan preventive measures
    - **safety monitoring:** implementing AI-powered monitoring systems to monitor compliance with safety requirements and identify potential hazards in real time
    - **energy optimization:** using AI algorithms to analyses patterns in energy consumption and operational data and recommend adjustments to optimize energy use and reduce costs /38/ /39/.
- **Applications used for control and monitoring:** here, AI applications begin to take more direct control of certain processes while still operating under human supervision. These applications are characterized by their ability to perform specific tasks autonomously based on predefined rules combined with machine learning. Human intervention only occurs when undesirable conditions occur and in connection with strategic decisions. Examples include:
    - **automated drilling operations:** AI systems that control drilling equipment adjust parameters in real time for optimal performance, subject to the human oversight by the operators /23/ /26/ /28/ /29/ /30/
    - **intelligent control systems:** AI-based control systems that handle the operation of valves, pumps, and other critical infrastructure components improve efficiency and reduce the risk of human error /22/.

- **Autonomous applications:** at the extreme end of the spectrum are fully autonomous AI applications, which operate independently of human intervention. While this level of autonomy is not yet widespread in the offshore industry, it represents a long-term goal for areas where human presence is particularly risky, or where AI can perform significantly better than human capabilities. Example of an area where AI can be expected to be introduced is:
    - **Autonomous Underwater Vehicles (AUVs) for inspection:** the AUV is expected to be equipped with AI to autonomously navigate and inspect the condition of underwater infrastructure without direct human control /40/ /41/. Should such a vehicle collide, the potential for damage is in many cases low. The threshold for introducing AI in the autonomous navigation of such vehicles is therefore expected to be correspondingly low.

4.5 Examples of risks associated with the use of AI

Typical reasons why AI produces unwanted results can be: deterioration of models, bias in data, bias in information models, poor data quality, and a lack of good data for training the algorithms. A lack of data from rare events can be an example of the latter. See chapter 5.6 for more details on typical causes.

The extent to which information produced with the help of AI can lead to major accidents depends on what the information is used for, the extent to which it is possible to detect the problem independently, and whether there is enough time to both detect the problem and implement measures. See also chapter 4.3 regarding the use of barriers.

Some examples where the use of AI can introduce risks are given below.

- AI is expected to be used in connection with predictive maintenance and maintenance planning. If AI is used to decide when maintenance should be performed, and this may be less frequently than what has been the norm so far, there is a risk that maintenance that should have been carried out will not be carried out, with a consequent increased risk of accidents.
- The use of advisory applications based on the continuous collection and analysis of data from different types of systems can also be used as an argument to increase test intervals, which can give rise to a risk that dangerous fault situations will be detected later than would be the case with more frequent test intervals.
- AI is expected to be used both in control systems and in applications that provide advice on what should be done in specific operations. Examples are applications and systems used in connection with lifting operations and drilling operations. There is a risk that information produced by AI may be incorrect or inappropriate for specific operational scenarios. This can be difficult for the operator to detect and can potentially cause problems in operation and lead to relevant emergency functions not being activated.
- AI is expected to be used in connection with the specification of requirements for, and design, implementation, verification, and validation of, systems. This will be relevant both for systems that use AI in their algorithms and systems that do not. This is in many cases expected to increase the quality of the processes - for example the use of AI can enable more extensive testing of systems. However, there is a danger that suppliers may come to rely too much on the tools, so that verification and validation processes involving humans become too weak.
- AI is also expected to be used to produce many different types of documentation, such as operational procedures, and here too there is a risk that verification and validation processes involving humans may become too weak.

5 METHODS TO ENSURE ROBUST SOLUTIONS

5.1 Technology qualification

Section 9 of the Facilities Regulations, "Qualification and use of new technology and new methods", states that:

> *Where the petroleum activities entail use of new technology or new methods, criteria shall be drawn up for development, testing and use so that the requirements for health, safety and the environment are fulfilled.*

The guidelines to this section 9 of the Facilities Regulations state the following:

> *In order to meet the requirement for a method for qualification of new technology, DNV-RP-A203 and "Oil & Gas UK Guidelines on Qualification of Materials for the Abandonment of Wells, issue 2", can be used.*

For qualification of AI, it will be DNV-RP-A203 that is most relevant, and the qualification process it recommends is shown in Figure 5-1.

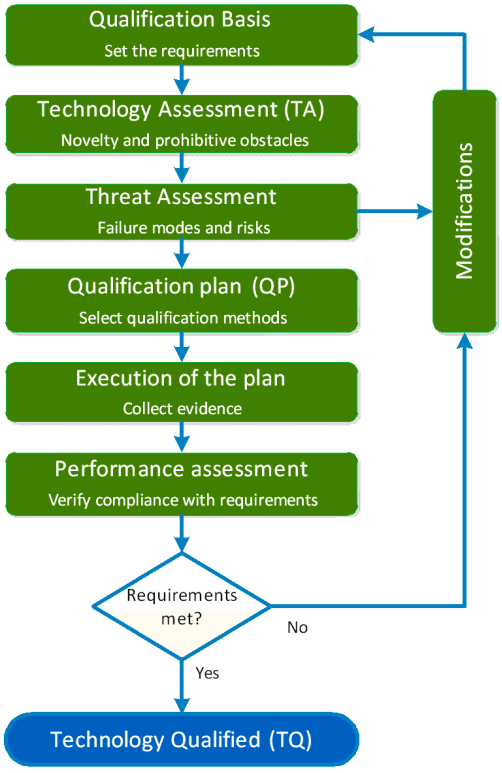

Figure 5-1 Overall TQ process using DNV-RP-A203.

The process is applicable to all types of technology, but for an AI-relevant project it will not be sufficient to use this guide alone for two reasons. The most important one is that DNV-RP-A203 is oriented towards the qualification of physical components and thus has limited guidance for the qualification of software-based functions. It will therefore be relevant to rely on other standards and guidelines identified in this knowledge overview, see the summary given in chapter 6.3.

The second reason is that the introduction of AI may lead to changes in how an activity is carried out and/or trigger a need for a change in strategy, for example in terms of which independent barriers that it is necessary to have in place. A demonstration of safety may therefore require changes beyond ensuring that the new technology performs as required in isolation.

For example, it is conceivable that the introduction of AI in a control function or an advisory function will require a new barrier strategy that switches from human activation of an independent safety function to fully automated

activation. This means that qualification of an AI solution can potentially trigger a need for a more advanced safety system with different types of sensors and a more advanced logic than what is required for manual activation.

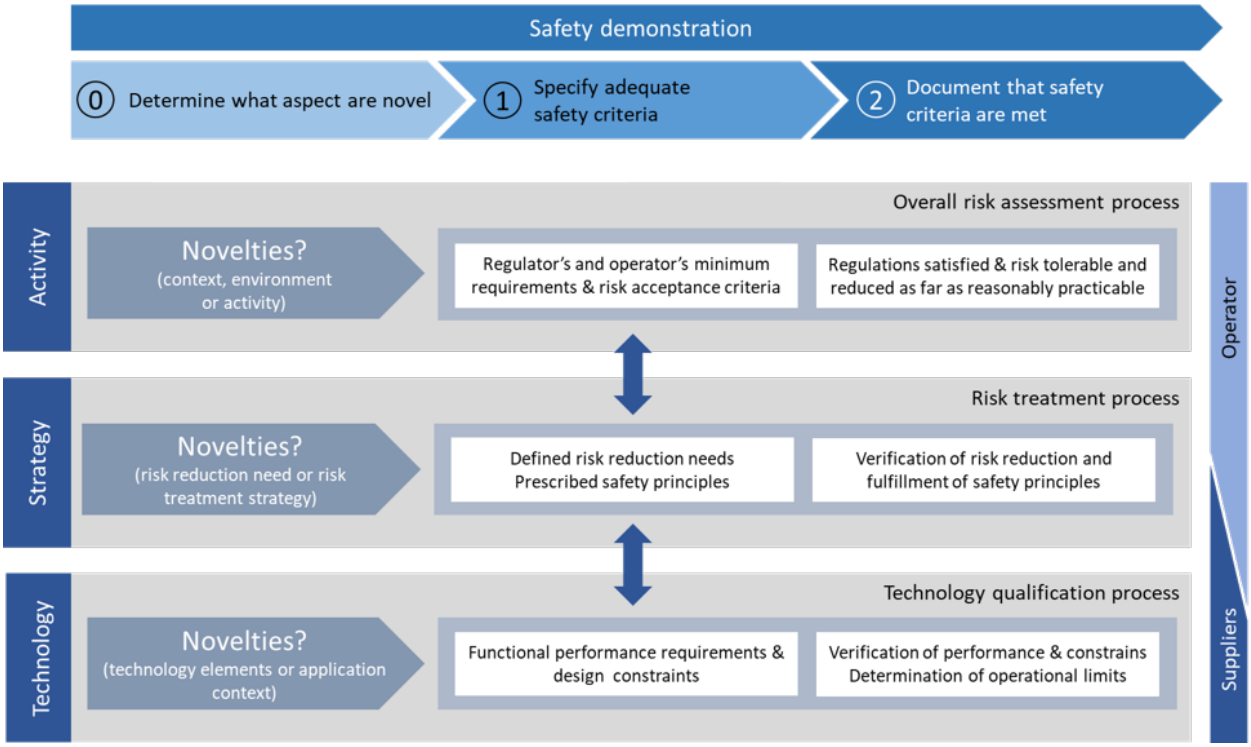

Figure 5-2 The introduction of new technology can lead to changes on several levels

Figure 5-2 illustrates that demonstrating safety in connection with the introduction of new technology may entail a need for changes that go beyond the technology qualification itself. This figure is taken from Safety 4.0, which was a Joint Industry Project (JIP) that looked at how to introduce new subsea technology in a safe way. Four oil companies, three suppliers, the universities of Trondheim and Stavanger, and DNV participated in this project.

The reports from Safety 4.0 were not prepared with AI in mind but compared to DNV-RP-A203 they have a greater focus on the system perspective described in chapter 4.2, and provide more guidance on managing software-related risks. The strategies currently used for this purpose are also relevant when it comes to the safe use of AI - see chapter 5.2 below for an overview of such strategies. The results of the Safety 4.0 project have also been used in the preparation of DNV-RP-0671 Assurance of AI-enabled systems /19/.

5.2 Management of software-related risk in existing systems

Weaknesses that can cause unwanted conditions to occur in software can be introduced at many different points in a system's lifecycle. This applies regardless of whether AI is being used or not.

Weaknesses can be introduced already when one is specifying requirements for and designing systems. This can happen because dependencies between the technical, operational, human, and organizational components are overlooked, because specifications are developed based on insufficient understanding of physical processes and the operational environment, or because possible unwanted input from other systems are not considered.

Weaknesses can also be introduced at the software design and implementation level, for example through unforeseen dependencies in the internal data flow or the unsafe use of the programming language. The software may also be insufficiently robust against degradation in the hardware platform on which it is being executed. Thus, unwanted conditions that are caused by degraded but still operational hardware can appear to be software related.

The software weaknesses will either be present in a system from day one in operation or be introduced through modifications. They will typically be hidden until specific circumstances occur, and there is currently no agreed method that allows calculation of the probability for:

- serious weaknesses being present in the software, and if so, the number of such weaknesses
- the values, combinations, and/or sequences of input data which are necessary to uncover these weaknesses actually occurring during operation.

Furthermore, when looking at the set of external interfaces that are relevant to a software function, the number of possible combinations and sequences of input data can often be extremely high. This means that no matter what type of technology stack being used, it is usually not possible to test the software exhaustively.

The test problem is usually further enhanced by the internal design of the software. One example is that the number of possible paths through the application software will often be virtually infinite due to different forms of feedback, for instance in state machines. Another example is that the complexity of the data flow within software can hide dependencies between different parts of the software, which is one of the reasons why unwanted side effects of modifications are quite common.

The fact that exhaustive testing is not possible, means that testing cannot demonstrate the absence of weaknesses in software. Testing is a very important way to reduce the number of weaknesses, but the fact that all tests developed for safety-critical software have been passed, is not sufficient evidence of safety. This generally makes it very challenging to demonstrate that risks associated with software is sufficiently low, if there are no independent barriers that can prevent a problem from escalating into a dangerous incident.

The large number of possible inputs also means it is often difficult to know whether the requirement specification for software is complete and fit for purpose because it is hard to foresee what kind of scenarios the software may be exposed to during the operational phase. This also means it can be uncertain whether mathematical algorithms to be implemented in the software are fit for purpose under all circumstances. Regarding the latter, introduction of AI may introduce further uncertainty.

Operational experience is often used as an argument for the software being safe to use. This is because the number of defects will typically be reduced over time if there have been no other types of modifications, than removal of defects. However, it will be difficult to assess how much risk reduction this provides, as combinations/sequences of input data that the software has so far never been exposed to may occur in the future. This means that even if software has worked for years without creating any dangerous situations, it cannot be ruled out that dangerous situations may occur in the future.

The introduction of AI-enabled software functions may further exacerbate the challenges described above, but in many cases, it will not fundamentally change the situation. The fact that AI causes systems to end up in unwanted states will often only be a special case of other software-related problems also creating unwanted states, so that the strategies used to deal with this problem will also be relevant to AI. This is discussed in the next chapter.

5.2.1 Existing strategies for managing software-related risks

From a high-level perspective, the strategies for managing software-related risks linked to health, safety, and the environment can be described as follows:

1. **Use of independent control functions.** There are many types of operations where it is possible to stop an individual control function, if necessary, but where a total shutdown/stoppage of the process being controlled does not represent a safe state. Examples in oil and gas activities are dynamic positioning systems and ballasting systems used on offshore drilling rigs. For such types of operations, the use of independent control

functions is usually necessary to maintain a safe state if the control functions that are normally being used end up in an undesirable state. Two factors that are important to consider when using this strategy are discussed below:

a. **Manual activation of independent control function.** In many situations, there are no independent functions that are automatically able to detect all forms of unwanted conditions regardless of the cause. In that event, the operator's ability to understand that an undesirable condition has occurred based on the total amount of available information will be crucial for safety, since in such cases the independent control function must be activated manually. The extent to which an independent control function represents a realistic alternative will also depend on whether the operator is able to activate it in time, and, if it has to be operated manually, whether this will be possible under all operational conditions.

b. **Independence between advisory applications and control functions.** In some operations, advisory applications are used as input data to operators' decision-making. AI used in such advisory applications may affect how operators use control functions to control a process, and undesirable results from the advisory application may therefore lead to the controlled process ending up in an unwanted state. The detection of the condition will typically take place through the control function's sensors, and the control function will also be used to bring the process to a safe state. In such cases, the advisory application and control functions will be technically independent of each other, and the control function may be part of an independent barrier. However, if there are no alarms from the control functions, detection that the controlled process has ended up in an unwanted state will depend on the operator having a very good understanding of the process being controlled.

2. **Independent safety functions** having a higher authority than the control functions are used to shut down the controlled process when this is necessary to maintain a safe state. See also "active safety functions" as described in Havtil's guidelines to section 8 of the Facilities Regulations /4/. Within the oil and gas industry, such safety functions are typically realized by means of standardized hardware and software components certified for use in safety-critical systems in accordance with the requirements of IEC 61508 /56/. There are two variants of this strategy, and these are significantly different when it comes to how demanding it is to demonstrate that the solution is safe.

   a. **Automated activation of safety functions.** In many systems, the safety functions are fully automated so that they are triggered based on readings from their own sensors that measure parameters such as pressure, temperature, loads, etc. If the value of such parameters is sufficient to describe the safe operating envelope of the controlled process, it may be possible to argue that AI in the control function will not adversely affect safety since the safety function will be activated if the AI algorithm drives any of the parameters out of a safe area. However, to avoid unnecessary shutdowns, it is important that the AI algorithm is trained not to challenge the limits that will trigger the safety functions.

   b. **Manual activation of safety function.** If there are any unwanted conditions in the primary control function that cannot be detected by independent sensor readings as described above, activation of the safety function may depend on the operator detecting that there is something wrong based on the total amount of information available. If that is the case, the picture becomes much more complex, since this raises a number of questions related to situation awareness and human factors. In such a case, it will typically be necessary to carry out a detailed risk analysis of that control function to identify relevant undesirable conditions and how these can be observed by humans if there are no alarms from that control function. It must also be considered whether the operator can realistically activate the safety function before it is too late. In such a setting, the introduction of an AI algorithm in the control

function can further challenge the operator's situation awareness and thus make demonstrating safety even more difficult.

3. **Operational restrictions** are used to reduce the worst effects if undesirable conditions in software escalate into a dangerous incident. In many cases, this is the easiest way to reduce the risk of harm to humans. A typical example is where people are not allowed to be near dangerous zones when an operation is taking place. This way of managing risk will also be highly relevant when introducing AI.

4. **Control functions developed to a high integrity level.** When this strategy must be used, the use of independent control functions and operational restrictions may also be relevant strategies, but these will not be able to reduce the risk sufficiently if the worst possible undesirable condition in the control function occurs. Thus, the control function must be developed to a high integrity level, and it is the supplier's safety management system, including the quality management systems, that govern the development of the system, hardware, and software, that contribute to the necessary risk reduction. IEC 61508 refers to such functions as "continuous safety functions".

The fourth approach is typical for critical systems in the aviation, automotive, railway, and a few other industries. It is very demanding in terms of both knowledge and resources, which means that suppliers in these industries will often be dependent on spreading their costs across many equal systems. Related to this, it is important to be aware that if AI is used for safety-critical systems, for example in the automotive industry, the algorithm will typically be implemented or integrated by organizations that have expertise in developing complex control functions to a high level of integrity. It is also important to be aware that applications included in such functions run on certified hardware and software. For example, operating systems will typically be a variant of a real-time operating system certified for use in safety-critical applications. VxWorks and QNX are two of many examples of real-time operating systems that have such certified variants. Since very few Norwegian companies supply safety-critical control systems for aircraft, railways, or cars, few people in Norway have experience with this strategy, which is by far the most demanding of the four discussed in this chapter.

In contrast, Norway has a large supplier industry in the oil & gas and maritime sectors, which are industries that are utilizing combinations of the first three strategies. This means that the approval assumes that control functions will fail from time to time, and that operational restrictions and/or independent safety functions and/or independent control functions, are therefore necessary to manage the associated risks. This philosophy is also recognizable from Havtil's barrier memo /7/.

In many cases, using one or more of the first three strategies will also be a far easier way to manage AI-related risk than using the last one, where it must be demonstrated that an AI solution is safe in isolation. However, in those cases where barriers need to be activated manually, there may be a lack of independent detection mechanisms. See 1a, 1b, and 2b above. The use of AI can potentially make it even more difficult for the operator to make the right decision, and more difficult to take control of the process manually in situations where this is relevant.

Thus, an increased degree of automation means that control functions that are not currently considered to be safety-critical are becoming more and more critical, which means that we are moving towards a grey area where one as in the automotive, railway, and aviation industries, can observe critically complex functions that must continuously work as intended to maintain a safe state. This challenges the approving process for safety-critical systems.

It is also a challenge that, as long as a function is not defined as a safety function and suppliers do not need to demonstrate a safety integrity level (SIL) or similar, suppliers are quite free to choose hardware and other types of off-the-shelf products such as CPU cards, operating systems, communication protocols, etc. They are also quite free to choose the methodology for development, verification, and validation beyond testing.

5.3 Risk acceptance criteria and how they can affect risk assessments related to AI

Below is an example of a risk matrix which contains acceptance criteria related to combinations of the probability of hazardous incidents and their severity. Risk classified as high will typically be considered unacceptable, while medium risk is often managed based on the so-called as low as reasonably practicable (ALARP) principle. The example matrix is not particularly strict, but it is still challenging to demonstrate sufficiently low probabilities for dangerous incidents that can have serious or catastrophic consequences.

Table 5-1 Example of a risk matrix with acceptance criteria.

Probablity of dangerous event per year	P >= x/year	None	Negligible	Minor	Significant	Severe	Catastrophic
Frequent	1.E+00	Low	Medium	High	High	High	High
Probable	1.E-01	Low	Low	Medium	High	High	High
Occasional	1.E-02	Low	Low	Medium	Medium	High	High
Remote	1.E-03	Low	Low	Low	Medium	Medium	High
Very Remote	1.E-04	Low	Low	Low	Low	Medium	Medium
Improbable	1.E-05	Low	Low	Low	Low	Low	Medium

In the oil and gas sector, it is typically operators that are responsible for defining these types of risk acceptance criteria. Havtil's regulations require such criteria to be in place and actively used, but do not define the limits for what is considered an acceptable risk.

Havtil's regulations are primarily function-oriented, which means they focus on achieving the desired outcome without prescribing precise technical solutions. However, there are a number of minimum technical requirements within the regulations, within Havtil's guidelines to the regulations, and in the standards and guidance referred to in the regulations. This means that even though different owners may have different risk-acceptance criteria, this does not mean that the general level of safety when it comes to standard technology used in the petroleum industry will be correspondingly different. However, different risk-acceptance criteria among owners may be a factor that must be considered when new technology such as AI is introduced, since no AI-related guidelines specific to the petroleum industry have been created so far.

For safety-critical functions where a “Safety Integrity Level” (SIL) or similar must be demonstrated, the probability of software failures in the safety functions is not calculated. The safety-standards prescribe so strict requirements for the processes relating to risk analysis, specification, design, implementation, verification, and validation that the probability of dangerous undetected software failure is negligible compared to the calculated probability of dangerous undetected random errors in physical components. This simplification can obviously be criticized but, in the absence of something better, it is typical of safety standards across industries - see, for example, IEC 61508 /56/. The strict requirements apply not only to the project-specific application, but also to all software that includes operating systems, communication protocols, drivers, etc. The requirements become stricter the higher the integrity that is to be demonstrated and go far beyond what is common for software in other contexts. This means that if you want to use AI in a safety function, you must be able to argue that the requirements applicable to software in relevant safety standards have been met, which will be very demanding.

In the petroleum sector, such strict integrity requirements are imposed only on independent barriers, and not on the functions that normally control the processes that the barriers protect. The threshold for using AI in control functions may thus be lower than the threshold for using AI in dedicated safety functions.

If, one during risk analysis of a control function, is using a risk matrix as shown above, it must be estimated how often this function may end up in a dangerous condition. The combination of probability and the worst possible consequence if the dangerous condition should lead to an incident will determine whether the risk is classified as high, medium, or low.

Such an estimate is difficult, both because there is no agreed method for calculating the probability of unwanted conditions in software, and because the assumption that very strict requirements for development methodology mean that this probability can be neglected will only be relevant for functions developed in accordance with safety standards. The use of AI in control functions will typically make it even more difficult to come up with such an estimate.

It is nevertheless important to be aware that the safety standards identified in Havtil's guidelines to section 5 of the Management Regulations impose some limitations related to the estimated probability of dangerous unwanted conditions, and that these may also be relevant for functions based on AI.

One of the limitations applies to functions that control a safety-critical process and have not been developed into a SIL or similar. For such functions, it must be assumed that a dangerous undetected fault will occur more often than every 10 years in operation, and this means that only the categories "Frequent" or "Probable" in Table 5-1 above can be selected. To claim an even lower probability of an undetected dangerous fault, the control function must either be developed into a SIL or similar, as is done in the aviation, railway, and automotive industries, or safety must be ensured by independent barriers, such as one or more independent safety functions and/or other types of barriers.

Havtil's regulations means that it is the barrier strategy that is relevant in the petroleum industry, and here it is worth noting that, for a barrier that requires human activation, the standards recommend setting the risk reduction factor (RRF) at a maximum of 10. Consequently, a barrier that is dependent on the operator will only reduce the probability estimate by a single level in the risk matrix above, which may result in too little risk reduction if the consequences of an event are high and there are no other independent barriers available.

5.4 Use of AI in safety functions

The introduction of AI in safety functions, where it is necessary to demonstrate a Safety Integrity Level (SIL) or similar, entails a considerable burden of proof. Safety functions in the petroleum industry are included in barriers and have a relatively simple logic. This means the extra complexity that the use of AI entails cannot easily be justified. For this reason, this report does not focus on the introduction of AI in such systems. However, it cannot be ruled out that AI-based components may be introduced in the longer term. For example, the AI-based activation of safety functions can be envisaged coming as an add-on, where today there is only human activation.

For further details on this topic, refer to ISO/IEC TR 5469 «Artificial intelligence—Functional safety and AI systems» /52/.

5.5 Cybersecurity

AI and cybersecurity can manifest themselves in different ways. Hackers can use AI for cyber-attacks, making the threats more sophisticated and effective (see 5.5.1). At the same time, AI can be used to defend against cyber-attacks by quickly identifying and reacting to abnormal activities (see 5.5.2). AI can also be the target of cyber-attacks, for example by manipulating training data for the AI algorithms or data poisoning (see 5.5.3). In addition, AI can support cybersecurity in system development by identifying vulnerabilities in the code and improving security measures (see 5.5.4).

5.5.1 AI as a threat

Hackers can exploit AI to carry out more sophisticated and effective cyber-attacks. This creates a few new cyber-security challenges, including:

- AI can enable hackers to automate attacks and execute them at scale

- AI can be used to create advanced phishing attacks that are more difficult to detect because AI can analyse large amounts of data to create more convincing scam messages that are tailored to individual goals
- hackers can use AI to develop new evasive techniques that bypass traditional security systems - for example, AI can be used to test defence systems to identify and exploit vulnerabilities.

An AI-based approach can also learn and adapt in a way that surpasses human attackers. This can improve the ability to carry out all these options.

5.5.2 AI as defence

AI can be used to improve cybersecurity measures in several ways. AI's ability to process large data sets quickly makes it possible to identify abnormal activities that may indicate a cybersecurity threat at a much earlier stage than with traditional methods. Through continuous learning mechanisms, AI solutions can also adapt to new and evolving threats faster than static, rule-based systems.

Furthermore, AI-powered security protocols can implement nuanced and dynamic responses to detected threats and thus effectively mitigate potential consequences. These adaptable measures can include:

- AI systems can automatically identify and classify emerging threats through comparison with a wide range of known patterns and anomalies
- by using advanced analytics and predictive modelling, AI can predict future attacks based on historical data and activity trends
- AI can develop automated defence tactics that adapt to enemy attacks in real time, minimizing damage and recovery time.

5.5.3 Attacks on AI

AI systems can also open new attack surfaces. For example, AI algorithms can prove vulnerable to manipulation. Machine-learning systems' reliance on data presents opportunities for different types of data-manipulation attacks where manipulated data can mislead AI systems.

In addition, AI systems will be vulnerable to the same types of deceptive actions as other software, including hacking, computer viruses, communication jamming, and sabotage by people with physical access. This generalized vulnerability underscores the need for comprehensive security measures to protect AI systems at all levels.

Examples of attacks on AI are:

- Manipulation of training data: hackers can manipulate the training data that AI algorithms rely on, thereby affecting the performance and behaviour of AI systems. This can render AI systems inefficient or cause them to perform malicious actions.
- Data poisoning: by contaminating the training dataset, attackers can ensure that AI models learn erroneous patterns, which can lead to erroneous decisions and actions.
- Model inversion attacks: attackers can reconstruct sensitive training data by analysing the model's responses, which can compromise the privacy and security of the data.
- Adversarial attacks: by making subtle changes to the input data, attackers can mislead the AI model into making incorrect classifications or decisions.

5.5.4 AI that supports cyber-security in system development

AI can assist developers and security analysts in several ways. AI tools can be used for automatic analysis of source code to identify vulnerabilities, including finding bugs that could be exploited by hackers or vulnerabilities that were not obvious during manual coding. AI systems can also monitor software for bugs and automatically log these in bug-tracking systems. This improves the efficiency of developer teams and reduces the time it takes to resolve issues.

AI can be used to continuously scan systems and networks for existing and new vulnerabilities and can suggest corrective measures to improve security. AI-based solutions can also be used for penetration testing (pen-testing) to mimic cyber-attacks and identify security holes before they are exploited by actual attackers.

5.6 Examples of risk factors specifically related to AI

This chapter looks at some risk factors that are specific to AI. It is important to take these into account in connection with risk analyses, assessing to what extent an operator or independent software-based function will be able to detect that results produced by AI are incorrect or inappropriate.

They are also important to consider when designing AI solutions that are as robust as possible, and they also illustrate that AI-based systems will typically require more follow-up in the operational phase than systems that do not use AI.

In addition to the factors mentioned in this chapter, there are a number of risks discussed in other parts of the report:

- Chapters 5.2 and 5.3 look at the risks associated with the fact that AI is usually implemented in software that is not developed to a high level of integrity, which creates a need for independent detection mechanisms.
- Chapter 5.5.3 looks at how AI can create new cyber-attack surfaces.
- Chapter 5.7 looks at the interaction between humans and AI and discusses, among other things, risks related to a lack of explainability and transparency, and the operator's ability to supervise the AI-based system.
- Chapter 6.2 provides a brief description of the EU AI act, and addresses the risks associated with an AI system that can be used for a different and more critical purpose than the one for which it was originally designed.

5.6.1 Model deterioration

Model deterioration occurs when the conditions under which an AI model was trained gradually change over time, or when unexpected variables are introduced into the operational environment. Deteriorated models can lead to unreliable predictions or recommendations, which in turn can have direct implications for operational safety and efficiency.

Causes of model deterioration can include:

- The operational conditions can change significantly over time. This can be caused by physical changes in equipment or resources, changes in staff skills, or modifications to operating procedures.
- The AI model can be exposed to new or unexpected data it was not trained on, which can reduce its accuracy or applicability.
- New technological advances may also introduce variables that were not considered during the initial model training phase.

Model deterioration can be prevented and managed in different ways:

- Periodic evaluation of the model's performance considering new data and changing conditions can provide early identification of deterioration. This should be done as part of the regular maintenance where, if necessary, new training of the model is run.
- Implementing adaptive learning systems where the model is continuously adjusted based on new information and feedback can help counteract the effects of model deterioration.

- Increasing the understanding of how models generate data can facilitate the identification of when and why a model begins to deteriorate.
- Automated rule-based assessments of the model's robustness against varied data and conditions can also play a role in anticipating potential weaknesses.

It is important that the operating organization has the knowledge and tools to recognize signs of model deterioration. This includes understanding model uncertainty, predictive reliability, and the ability to override AI-based decisions when necessary.

5.6.2 Bias in data and information models

Data and model biases can materialize in several ways, and systematic efforts are required to identify and manage these biases effectively. These biases include:

- Feedback loop bias: potential positive feedback loops where the model's output affects subsequent data inputs. Implementing control and balance measures in model outputs can prevent the snowball effect from initial biases.
- Historical bias: it is important to recognize, and correct biases embedded in historical data, to ensure that AI models do not perpetuate past inaccuracies.
- Inductive bias: this refers to a collection of assumptions, expressed either directly or implicitly, on which a learning algorithm bases the induction process. Incorporating domain knowledge into AI models help counteract inductive bias and provides a structured way to introduce useful assumptions.
- Measurement bias: verifying data inputs helps mitigate bias stemming from erroneous measurements or data collection methods.
- Representation bias: ensuring diversity in data collection processes to protect against bias that could arise from the over- or under-representation of certain patterns or groups.

Counteracting these biases requires a continuous process that needs to be incorporated into a data-governance framework.

5.6.3 Deterioration in data quality

Data forms the basis for building and training all AI models. The quality of the data therefore plays a critical role in the model's performance, reliability, and resiliency. Data quality involves aspects such as adequacy, completeness, accuracy, and relevance, as well as the absence of bias.

The challenge in many AI projects is not necessarily a lack of data, but the availability of high-quality data that is representative of the real operational conditions. There is frequently a large amount of data for the nominal operation of a system, but a lack of data regarding accident states, fault states, or abnormal operating states that there is often an interest in predicting or detecting with the help of AI.

Some examples of data-quality issues (see ISO 8000) are as follows:

- Incomplete data
- Noise in data
- Incorrect metadata or incorrect data classification
- Inconsistent or unreliable data collection
- Incorrect or irrelevant data
- Data aging
- Lack of representativeness
- Hidden biases in data
- Data fragmentation

A lack of relevant real-life data means that, in some cases, both real data and synthetic data produced using AI are used to train AI. If this is done iteratively, more and more of the training will be based on synthetic data and there is a risk that the results will increasingly converge and become useless, which is called model collapse. However, using synthetic datasets that are designed to mimic statistical properties typically found in real-world datasets can also be used to counter the problem /71/.

5.6.4 Overfitting

This risk is about over-adaptation to training data, where an AI model adapts too well to training data that includes irrelevant noise and details. This means that the model performs well based on training data, but poorly based on new data.

5.7 Interaction between humans and artificial intelligence

Modern AI-enabled systems typically use probabilistic models and machine-learning algorithms that are trained on extensive datasets to identify patterns in data. While these systems are good at pattern recognition, their drawback is that the relationship between input and output parameters may be neither entirely understandable nor predictable to the user /72/ /73/. Since AI-based systems "don't know what's possible in the world", developing reliable and predictable systems remains a significant challenge /73/. Given these limitations, such systems are not currently sufficiently suited to independently handle new and complex situations that require careful oversight and management /74/. This means that, for the foreseeable future, human oversight is likely to remain essential for monitoring system performance, guiding operations, and ensuring that desired outcomes are achieved /75/. However, due to their probabilistic nature, these systems are more difficult to monitor and predict than "traditional" control systems /76/. As a result, operators may find it difficult to understand and assess these systems' behaviour /75/ /87/ and thus that their ability to monitor them is challenged.

In this context, trust is a relevant psychological concept to consider when designing for human-to-human interaction and automation /89/. Here, trust is defined as "the attitude that an agent will contribute to achieving an individual's goals in a situation characterized by uncertainty and vulnerability" /88/ (p. 51). Trust is important in the interaction between humans and AI systems as humans tend to use systems they trust and reject systems they do not trust. However, there may be too little or too much trust in the system. People can become overly critical of a system and thus choose not to use it at all ("disuse"), or they can become uncritical and trust the system too much ("misuse") /84/. Consequently, when designing for trust in AI-enabled systems, these must be built to be robust and reliable. Furthermore, these systems must be able to support human decision-making and allow for intervention when necessary. Therefore, efforts must be directed towards designing systems to support effective interaction between humans and automation. Figure 5-3 illustrates the aspects that should be considered to design and operate an advanced and robust system.

- **Human capacity** – assess the operator's cognitive load, correct situation awareness, and ability to respond to an undesirable situation.
- **Technology interface** – the user interface should be designed so that the operator receives correct and relevant information about the operation and state of the system.
- **Competence** – the quality of user training - with a focus on being able to use the system to perform the intended operation, monitor and understand the system, and perform correct actions in the case of fault conditions or unsafe situations.
- **Governance model** – a mature governance model for the development, operation, and maintenance of the system and follow-up of the users. This includes roles and responsibilities, approved procedures, and follow-up from management.

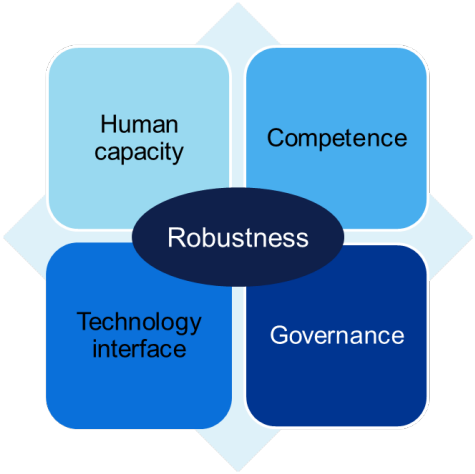

Figure 5-3 Aspects for a robust human-centred design

5.7.1 Interaction between humans and automation

The interaction between humans and automation has long been a research topic. One goal of automation systems is often to replace "unreliable" human behaviour with "reliable" automation. Unfortunately, there may be a discrepancy between the system designers' intention with the system and the implemented system, and it may also be the case that the system cannot be automated as much as desired. For example, a safety-critical system may lack the ability to detect certain undesirable conditions regardless of what has caused them, and thus automatic transition to a safe state will not always be possible. This leaves the human operator, tasked with overseeing the automated system, with (a potentially arbitrary set of) remaining tasks and little support for performing these /77/. Paradoxically, such (inappropriate) design and use of automation can exacerbate, rather than eliminate, human "unreliability" as humans are no longer an active part of the system's information processing and decision-making process.

People who are removed from the information-processing and decision-making loop may find it challenging to build and maintain situation awareness and end up "out of the loop" (OOTL) see Figure 5-4) /78/. Research has shown that this leads to a number of challenges to human performance and ability to adequately monitor. These include complacency (the assumption that "everything is fine") /79/ /80/, automation bias (the assumption that the system is likely to be correct) /81/, decreased alertness (due to depletion of mental resources), and abuse of the system (due to over- or under-reliance on the system) /84/. Finally, when automation fails, the operator may not be fully up to date with the current state of the system, resulting in an unreasonably high workload in an attempt to restore control /85/. This is especially relevant for systems with a high degree of automation or autonomy.

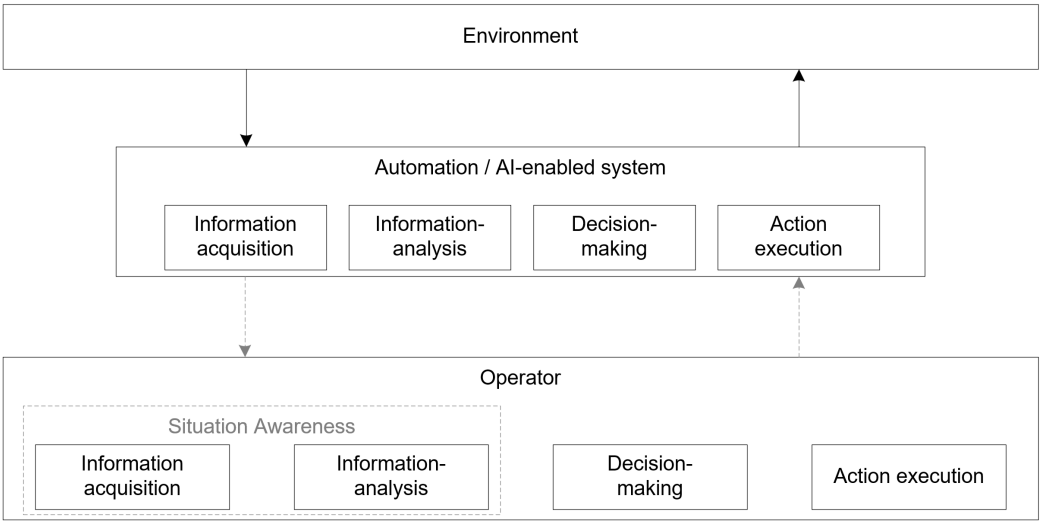

Figure 5-4 Inadequate insight into the system's information processing may lead to reduced situation awareness and the operator being "out of the loop" /78/

The above automation-related challenges are equally relevant to AI-based systems /111/. At the same time, there are new challenges that are specific to AI-based systems /86/. For example, it will be challenging to understand how a system works when it can change over time. Systems that receive frequent updates, or have the ability to learn over time, will be challenging as their behaviour will not be as predictable as that of traditional systems. Furthermore, decision-support systems can use large language models to develop and communicate plans. Given the realism of the language used in such systems, the formulation of these plans may seem convincing to humans. This can increase human propensity to agree with the system's suggestions rather than consider them critically. This is especially relevant if systems are perceived as reliable, which in turn leads users to believe these systems are more capable than they really are /86/. Moreover, users may lose focus and engage in other tasks /92/ /93/. Considering the widespread use of AI community-wide-based systems for low-risk applications, such as large language models available in standard Office applications, this can render the user desensitized to the system’s shortcomings. Consequently, the risks associated with these systems can inadvertently migrate from low-risk to high-risk applications.

These examples illustrate that people who are tasked with monitoring advanced and capable AI-based systems are expected – but possibly not able – to intervene when necessary. Consequently, operators often end up accepting the system's proposals, so that they become trained "button pushers" who lose their ability to perform their monitoring tasks adequately /94/. This means operators may not be able to adequately act as an independent safety barrier. To sum up, the interaction between humans and automation is an intricate problem – the more capable and reliable the system becomes, the less likely the operator is to take manual control of the system in a critical situation, and the greater the consequences of inaction /75/.

5.7.2 Strategies to support operator performance

A number of strategies have been developed to support interaction between humans and automation /75/ /95/ /96/ /97/. Since the OOTL problem is characterized by a reduced ability to detect system errors and perform tasks manually when automation fails, a key mitigation strategy is to re-involve humans in the system's decision-making process /78/. This means that, when designing for human monitoring of the performance of automated systems (including AI-based), consideration should be given to what operators are expected to do, how they are expected to do it, and what information they need for this. In this way, involvement in the task and the human decision-making process can be designed by the human automation team.

Human Centered Design (HCD) is an approach in which functions, tasks, and human needs are taken into account in the system design /95/ /98/. Unlike technology-centric design, HCD considers the user's tasks, context, and abilities and uses these to inform the system design, for example when designing human-machine interfaces. To achieve an HCD process, the following main activities should be performed (see Figure 5-5) /98/:

- the system's usage context must be specified and understood
- user requirements must be specified
- design solutions must be produced
- the solution must be evaluated and tested.

Through a series of iterations in which user requirements and design solutions are updated and evaluated, a design solution is produced that will meet the user requirements. Previous research and experience from using this process for new buildings and modifications on the Norwegian continental shelf have shown that the HCD process is an effective method for developing human-centered automated systems /99/ /100/ /101/ /102/. This means that system designers who are tasked with developing human-centred AI-based systems have an established process at their disposal to develop systems that support human monitoring.

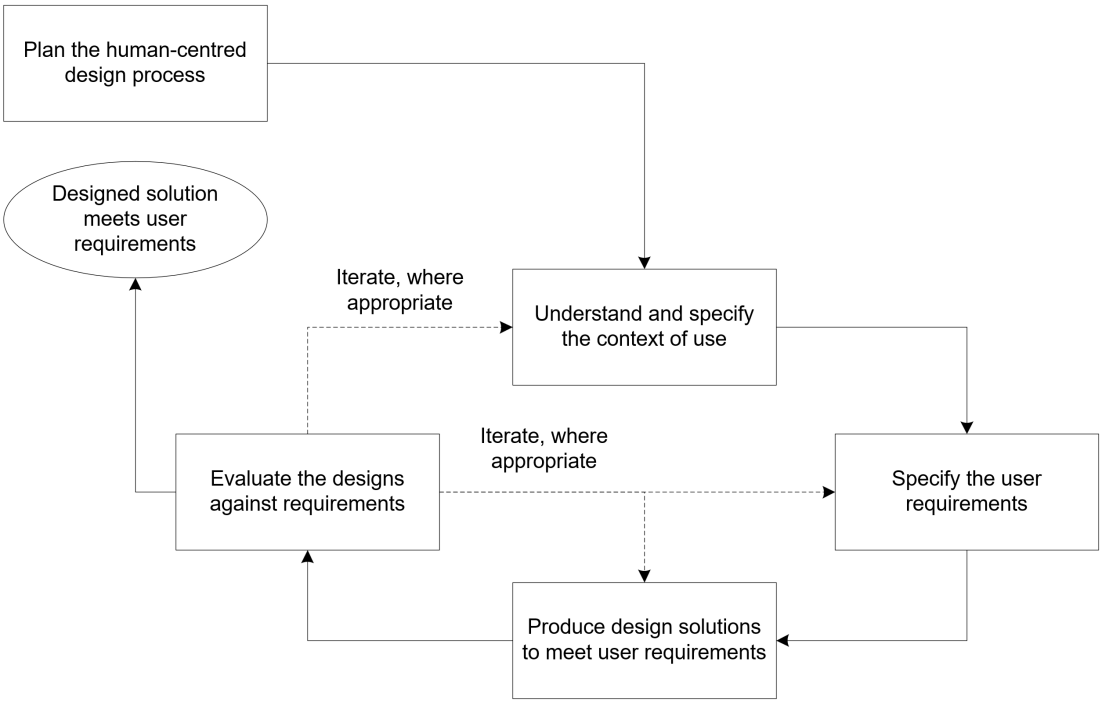

Figure 5-5 Human-centred design process /98/

When designing AI-based systems, emphasis should be placed on people's need to understand these, on minimizing their complexity, and on providing support for situation awareness. Here, adequate and appropriate feedback from the system is one of the key elements in helping people to create adequate mental models of the system /103/. However, what counts as "adequate and appropriate feedback" varies depending on the task and the context of the operation. For example, systems for time-critical operations have different requirements than systems where there is more time to assess available information and make better-founded decisions. This means that, in time-critical situations where operators need to make quick decisions, the system’s information base and interaction opportunities should be adapted to the operator's needs. This may involve presenting few yet essential parameters, using visual or audio cues, and providing concise description of suggested actions.

Applications that are not time-critical do not need to limit the amount of information to the same extent as time-critical applications, as the operator has more time to absorb information and consider possible measures. However, what is time-critical is highly dependent on the system's context and choices regarding decision-making authority, division of functions, tasks, and performance requirements distributed between the system and operator. In any case, system information should, as a minimum, correspond to the operator's decision-making strategies by presenting relevant information on a user interface that ensures the operator has the right situation awareness /104/. In addition, requirements should be set for the system's own ability to detect fault conditions or increased uncertainty. This means that, if the digital system is degraded in some way so that the information presented may be incorrect, then the operators must be warned so that they can take this into account and return the system to a safe state.

When an operator is tasked with monitoring and interacting with AI-enabled systems, the system's comprehensibility and predictability become important. Because these systems can change over time, it becomes increasingly difficult to train operators to maintain accurate mental models. Mental models are important mechanisms that humans use to understand how systems work and what information is considered necessary to interpret system behaviour and maintain proper situation awareness /91/. Therefore, explainability plays an important role in revealing how the system's algorithms work and the system's reliability, model performance, contextual information, risk explanations, prediction accuracy, and integrity /89/ /105/. Furthermore, transparency plays a role in building user trust in the system so that an appropriate dependency on the system is promoted and appropriate decisions about automation use can be made (see Figure 5-6) /106/ /107/ /108/ /109/.

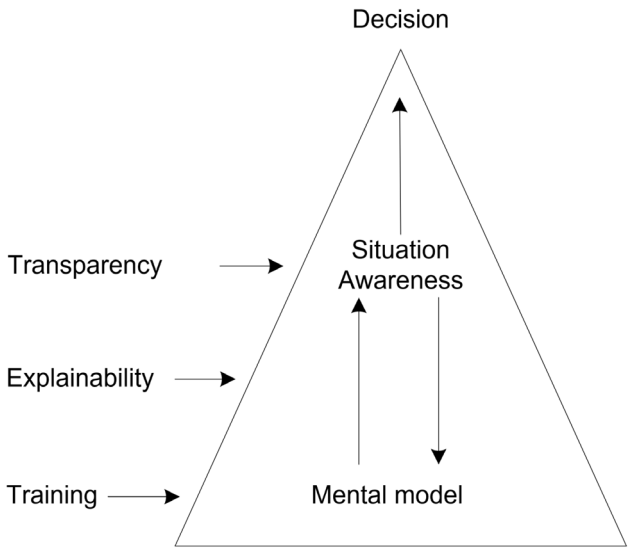

Figure 5-6 Explainability and transparency in a decision-making context /74/ /87/

When operators interact with systems that make recommendations or perform actions that have safety-critical implications, insight into the system's reasoning is essential for effective monitoring /109/. Therefore, such systems should be able to provide "explanations" and be "transparent" about their decisions and actions. In this context, explainability provides insights into the logic, process, factors or reasoning of the AI, that is the bases for the output from the AI system /74/, p.31). In other words, explainability helps the user to understand the logic used by the AI algorithm and by providing retrospective information that forms the basis for the decisions. In this way, the explanatory ability supports the operator's understanding of how the system works, when it will work, and when it will not work /87/.

Transparency "provides an understanding of the actions of the AI system as part of situation awareness" /74/, p.31. Transparency supports the operator by providing possible information about how the system works in real time and

planned actions in the near future. This means that transparency helps the operator to understand the system's situation awareness, decisions, and actions. In sum, explainability supports and maintains the mental models that underpin the operator's understanding of how the system works, while transparency directly supports the operator's situational understanding of the system in its task environment by providing current and relevant information in real time /87/ /102/ /110/.

Note that what has been described so far in this chapter only constitutes the first line of defence if an operation ends up in an unsafe state caused by AI. The safety philosophy in the petroleum industry is based on the use of independent barriers, and not on individual functions that are part of control systems or advisory applications being developed to provide a high level of integrity. This means that one cannot assume that there will always be an alarm if an unsafe situation occurs, and therefore there is a need to detect the unsafe state through information that has been produced independently of the AI-based system.

In some cases, such independent detection may take place before information from the AI system is used - for example, operators can in some cases compare results from an advisory system that uses AI with results from other advisory systems that do not use AI. In other cases, independent detection will only be possible based on the observation of data from the process being controlled.

Sometimes, independent barrier functions can be activated automatically based on the measurement of very specific process parameters, while at other times a human will have to make the decision about activation based on the total amount of information available. In the latter cases, the ability to understand that the process being controlled has ended up in an unsafe state may be important. Such an understanding requires either that the system controlling the process is independent of the AI-based system, and thus can raise an alarm based on the process state, or that the operator is able to draw conclusions based on different measurements of the process being controlled. If the operator is to be able to draw conclusions based on raw data from measurements, this requires deep insight into how the process being controlled is working. In most cases, such insight will be far more valuable than insight into how the AI algorithm works.

Regardless of the level of knowledge, it may be difficult to identify abnormal situations manually if the deviations from the normal process are relatively small, which may be a problem if one is also operating with small margins against unsafe process conditions. In addition, measurement data from the controlled process may not always represent independent information, since there may be opportunities for common cause failures in the system that controls the process, e.g. in software. See also chapters 4.3, 5.2, and 5.3 regarding the use of barriers and issues related to independent detection.

The industry should therefore explore the possibilities for further automation of independent barrier functions.

5.7.3 Towards Human-AI collaboration

There are potentially major benefits to be gained from combining people and AI-based systems, but the implementation of such solutions should be carefully managed. The development in AI is rapid and knowledge is constantly evolving. This also means that the research field on interaction between humans and AI systems is constantly evolving. Despite the efforts to create an overview of the current state of knowledge in the field, there are many uncertainties and unresolved issues /74/. Existing knowledge about human-factor methods is applicable and useful for AI-based systems, but there is an urgent need to further develop knowledge across a number of dimensions about how people and AI-based systems can effectively collaborate. Therefore, efforts should be directed towards ensuring that people are able to understand and predict the behaviour of the system and have methods to understand when to trust the system and when not to trust the system so they can make good decisions based on available information and at the same time exercise control over the system in a meaningful and timely manner /74/.

5.8 AI-generated code

AI-generated code encompasses the automated generation of software code using machine-learning models and algorithms. One way of working would be for developers to let AI create a draft that they then improve through manual coding, but it could also be relevant to use AI-generated code without modifying it in any way.

When it comes to verification and validation requirements for such code, it is natural to draw comparisons with how code that is automatically generated by models is handled. Such generation has for decades been common practice, also in the development of safety-critical systems, in various industries, such as aerospace and automotive. For safety-critical code, there are two models that can be followed:

1. The code is verified and validated as if it had been written by a human. If this is the case, AI-generated code is not handled differently than code written by a human.
2. One chooses to trust the generator and does not carry out all verification and validation (V&V) activities that would be performed for code created by humans. For example, manual inspections may focus on the model from which the code is generated, while manual code inspection and other implementation-oriented V&V activities may not be carried out.

If one follow model number two above, this will trigger requirements for qualification of the tool that generates the code for safety-critical systems. For highly critical code, it will probably be impossible to get such an AI-based tool qualified today.

Similarly, tools that automate parts of the V&V process, for example by generating executable sets of tests, will have to undergo qualification if what is produced is not verified and validated by a human.

Within aviation, there is a guideline for model-based development and verification of software. This describes which parts of the standard guidance are to be omitted and what is to be added in the case of model-based development and/or model-based verification.

When it comes to AI developed for use in safety-related systems in the oil and gas industry, it will be natural to follow a similar model. Since there is currently no regime for qualifying tools for software that is not subject to SIL requirements, the easiest way would be to carry out the same V&V activities that are carried out for manually generated code. However, the fact that the V&V requirements are not particularly strict may be a challenge. This is especially true for advisory systems.

5.9 AI-generated documents

AI must be expected to be used in connection with the production of a wide range of different types of documents. This applies to everything from work procedures to technical specifications and test procedures. As in the case of source code above, it will be important that personnel with relevant expertise carry out manual verification and validation of such documents.

6 REGULATORY AND STANDARDIZATION REQUIREMENTS

6.1 Havtil's regulations

Havtil has established five sets of regulations for oil and gas activities: *The Framework HSE Regulations* /1/ and *The Management Regulations* /2/ contain general principles for risk reduction and risk control. *The Activities Regulations* /3/, *Facilities Regulations* /4/, and *Technical and Operational Regulations* /5/ describe operating and design principles. As illustrated in Figure 6-1 below, the Framework HSE Regulations and Management Regulations are superior to the other three. This regulation applies to health, the work environment, and safety at onshore facilities.

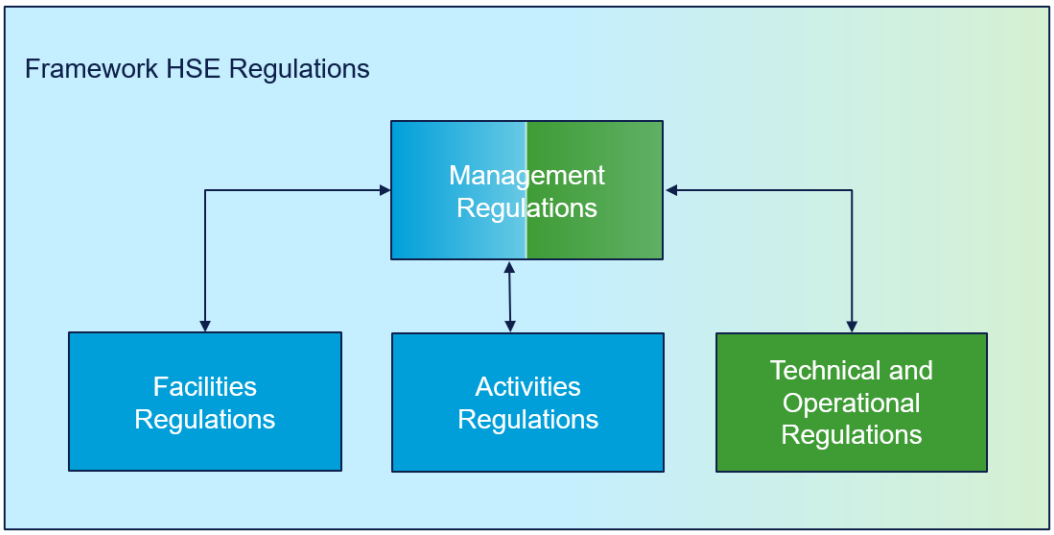

Figure 6-1 The structure of the regulations /6/

Separate guidelines for the regulations show how the regulations' provisions can be fulfilled. The regulations and guidelines must be read together to obtain the best possible understanding of how the regulatory requirement is to be met. In some areas, the guidelines refer to industry standards as a recommended way to meet the requirements of the regulations. The guidelines for the regulations are not legally binding, and the players can therefore choose other solutions.

If the responsible party chooses to use the recommended solution, it can normally be assumed that the regulations' requirements are met. If it chooses other solutions, such as other standards or company-specific procedures, the player must be able to document that the chosen solution is at least as good as, or better than, the recommended one.

It is the player's responsibility to follow the regulations, which are largely formulated as *functional requirements* with a focus on needs and desired results, rather than on prescription of specific solutions. This gives the players a significant amount of freedom, but it also requires an active approach where the players must define their own risk acceptance criteria and performance requirements.

A fundamental design principle is that the failure of a component or system or a single fault shall not lead to unacceptable consequences (Section 5 of the Facilities Regulations). How this is to be achieved is described in more detail in other parts of the regulations - for example that a system must enter or maintain a safe state if a subsystem fails (Sections 33 and 34 of the Facilities Regulations), the use of several barriers/redundancy (Sections 33 and 34 of the Facilities Regulations), the independence of the barriers (Section 5 of the Management Regulations, Sections 33 and 34 of the Facilities Regulations), and the ability to detect problems and know the status of the barrier (Section 26 of the Activities Regulations, see Section 34a of the Facilities Regulations).

In addition to the guidelines that are directly linked to the regulations, Havtil has published memoranda related to barrier management /7/ and risk management /8/ which provide further guidance regarding the paragraphs mentioned above.

> ***" Barriers supplement a safe and robust solution "***
>
> *Regardless of the efforts made to secure a safe and robust solution, failure, hazard and accident situations will occur. Barriers must then fulfil their functions in order to help handle such situations.*
>
> **From the Barrier Memorandum /7/**

Seen in the context of AI used in applications that are safety-related, the barrier memorandum text above means operators should do everything they can to make such an AI solution secure and robust but must also have effective independent barriers.

The risk-management memorandum particularly emphasizes that decisions must consider uncertainty, the precautionary principle, and learning. It uses the term risk-informed enterprise management to clarify that decisions should not be made based on risk analyses alone.

> ***"Taking account of uncertainty means systematically seeking out potential surprises»***
>
> *The conclusion is often drawn that an incident can be ignored because of its low probability. Such probability assessments can build on inaccurate or weak assumptions. "Taking account of uncertainty" means concentrating systematically on this problem and seeking out potential surprises. It is particularly important in this work to be aware of what is known in the organisation or in the industry beyond, but unknown to those making the assessment ("unknown knowns").*
>
> **From "Integrated and unified risk management in the petroleum industry" /8/**

As described in chapter 5.1, the regulations also require the qualification of new technology.

6.2 The EU regulation on artificial intelligence (EU AI Act)

The purpose of the EU regulation on AI (AI Act) /50/ is to contribute to the optimal functioning of the EU's internal market through harmonized legislation for the management of AI, as well as to contribute to a high level of protection against negative effects from the introduction of this type of technology.

The EU regulation sets strict requirements for systems that are classified as high-risk AI systems, and the rules that determine whether a system falls into this class can be found in Chapter III, Section I, Article 6 of the regulation. See the excerpt below.

1. *Irrespective of whether an AI system is placed on the market or put into service independently of the products referred to in points (a) and (b), that AI system shall be considered to be high-risk where both of the following conditions are fulfilled:*

   (a) the AI system is intended to be used as a safety component of a product, or the AI system is itself a product, covered by the Union harmonisation legislation listed in Annex I;

   (b) the product whose safety component pursuant to point (a) is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment, with a view to the placing on the market or the putting into service of that product pursuant to the Union harmonisation legislation listed in Annex I.

2. *In addition to the high-risk AI systems referred to in paragraph 1, AI systems referred to in Annex III shall be considered to be high-risk.*

Excerpt from Chapter III, Section I Article 6 of the EU Regulation on AI /50/

Item 1 of Article 6 refers to Annex I, which in turn refers to EU legislation relevant to the type-approval of specific types of products and systems. Several of these types of systems are used in the petroleum industry, for example systems that must be in accordance with the Machinery Directive /69/ or with the Directive for equipment used in potentially explosive atmospheres /70/.

Item 2 of Article 6 refers to Annex III, which identifies systems that are used in different areas of society and will end up in the high-risk category if they use AI. Several of these areas may be relevant to the petroleum industry, e.g. systems used by human resources departments may end up in the high-risk category if AI is being utilized.

The area considered most relevant to this report and mentioned in Annex III is AI systems used as safety components within critical infrastructure, since systems used for gas deliveries fall into this category, as shown below.

Critical infrastructure: AI systems intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity.

From Annex III of the EU Regulation on AI /50/

The focus on precisely these types of critical infrastructure is because the loss of deliveries can pose a danger to the population. This is reflected in the fact that the definition of a safety component below encompasses more than the usual definition of a safety system in the petroleum industry.

'safety component' means a component of a product or of an AI system which fulfils a safety function for that product or AI system, or the failure or malfunctioning of which endangers the health and safety of persons or property;

From Article 3 "Definitions" of the EU Regulation on AIAI Act /50/

DNV's understanding of the two definitions above is that both traditional safety systems and systems that can cause a loss of gas deliveries will be considered a high-risk AI system under the AI Regulation if they use AI.

Article 6 also contains some exceptions that will be relevant when classifying systems. These exceptions are systems that fall under the areas described in Annex III, but which are considered to be lower risk because they should: perform simple procedures, improve results from human activities, be in addition to human decision-making processes, or only be used for preparatory work. See the detailed description below.

These exceptions illustrate that systems used for advisory purposes, planning, and condition monitoring as described in Chapter 4.1 may end up in a grey area when it comes to classification according to the AI Act.

By derogation from paragraph 2, an AI system referred to in Annex III shall not be considered to be high-risk where it does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.

The first subparagraph shall apply where any of the following conditions is fulfilled:

(a) the AI system is intended to perform a narrow procedural task;

(b) the AI system is intended to improve the result of a previously completed human activity;

(c) the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or

(d) the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III.

Notwithstanding the first subparagraph, an AI system referred to in Annex III shall always be considered to be high-risk where the AI system performs profiling of natural persons.

Excerpt from Chapter III, Section I Article 6 of the EU Regulation on AI /50/

AI algorithms will typically be generic algorithms that can be used for many different purposes. However, the EU AI Act places great emphasis on AI systems being created for a specific purpose (intended purpose), and that everything from risk analyses to the training of algorithms must be adapted to this purpose.

There is a possibility that AI systems that were created for a specific purpose may eventually be used for another purpose. Chapter III, Section 3 Article 25.1.c. of the Act points out that if a type of AI-based system that is already in use is being used for a new purpose which means that it is now classified as a high-risk system, then all the Act's requirements for high-risk systems must be met.

Players that develop and operate high-risk systems must have risk management and quality-assurance systems in place, and must, among other things, meet regulatory requirements for testing, performance, data management, and documentation before the systems can be implemented. The requirements are extensive and also implies that the AI systems must be transparent, traceable, and explainable to ensure that decisions made by AI can be understood and verified by those who operate the systems.

Currently, individual players that want to use AI must specify and respond to the Act's requirements in their own management systems. Such operationalization of high-level requirements is usually labour-intensive, but it should be possible to reduce the workload on each individual organization if the industry joins forces to produce joint guidelines.

6.3 Relevant international standards

ISO/IEC 5338: Information technology – Artificial intelligence – AI system life cycle processes /51/ provides guidance on life cycle management of AI systems, with an emphasis on the systematic and structured approach required from inception to decommissioning. It requires rigorous quality assurance and project management processes to ensure that AI systems are reliable, secure, and efficient.

ISO/IEC 42001: Information technology – Artificial intelligence – Management system /55/ lays the foundation for establishing an AI management system and outlines requirements for management, planning, support, operation, evaluation, and improvement. ISO 42001 is built on the same template as ISO 9000, ISO 31000, ISO 27001, and others. It is possible to be certified according to it.

ISO/IEC TR 5469: Artificial intelligence – Functional safety and AI systems /52/ This is a technical report that aims to make it possible for developers of safety-related systems to use AI technologies in safety functions by providing a better understanding of: the properties that different AI technologies have, what the risk factors are, which methods

within the field of functional safety will be relevant, as well as possible limitations. The report has many links to IEC 61508.

ISO/IEC 22989: Information technology – Artificial intelligence – Concepts and terminology of artificial intelligence /53/ provides an overview of basic AI concepts and terminologies. ISO/IEC 22989 covers aspects such as explainability, robustness, risk management, and the life cycle of AI systems.

ISO/IEC 23894: Information technology – Artificial intelligence – Guidance on risk management /54/ focuses directly on risk management related to the application of AI systems. This standard provides instructions for identifying, assessing, and managing risks associated with the use of artificial intelligence in operational contexts. It emphasizes the importance of a continuous risk-assessment process to detect and mitigate potential threats to the safety and reliability of AI-integrated systems.

IEC 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems /56/ specifies the functional safety requirements for electrical, electronic, and programmable electronic systems used for safety-related purposes. The standard covers the entire life cycle of a system, from the early concept stage to decommissioning. It is a cross-industry standard that is applicable in many sectors where safety-critical systems are in use. Havtil's regulations refer to this standard in sections dealing with safety functions, and it is common for components used in safety functions in the petroleum industry to be certified in accordance with this standard.

IEC 61511: Functional safety – Safety instrumented systems for the process industry sector /57/ is derived from IEC 61508 and is specifically aimed at safety instrumented systems (SIS) used in process industries, such as the chemical, petrochemical, and oil and gas industries. The standard has an expectation that all software and hardware components used in a SIS have been developed in accordance with the relevant requirements of IEC 61508, which means that IEC 61511 can have a system focus. Havtil's regulations refer to this standard in sections that deal with safety functions

CEN-CLC/JTC 21: CEN and CENELEC have received a standardization request related to high-risk AI systems from the European Commission. In this regard, CEN-CLC/JTC 21 is now developing European standards. The European standardization organizations have until the end of April 2025 to finalize and publish these standards. The Commission will then assess and, when ready, approve the standards, which will be published in the Official Journal of the European Union. When published, the standards will provide a "presumption of compliance" with the EU AI Act for AI systems.

6.4 Other regulations and guidelines relevant for the prudent use of AI

There are several regulations and guidelines that have been developed to ensure safety when using AI. Below is a description of the most relevant ones.

- AMLAS (Assurance of Machine Learning for use in Autonomous Systems) /58/ is a methodology developed to integrate safety assurance into the development of machine-learning components. AMLAS covers several phases, including safety assurance, data governance, and model verification, and helps establish an evidence base for the safety of AI-components when integrated into autonomous systems.
- DNV-RP-0671 «Assurance of AI-enabled systems» /19/ is a recommended practice prepared by DNV that specifically focuses on the assessment of AI systems.
- The studies «Management System Support for Trustworthy Artificial Intelligence» /60/ and «Safe AI - how is this possible» /64 from the Fraunhofer Research Institute in Germany provide guidelines for the implementation of reliable AI systems. They emphasize transparency, verification, validation, and the need for human oversight in critical AI applications.

- NIST (National Institute of Standards and Technology) has developed guidelines for handling bias in AI systems /61/. These guidelines focus on identifying and limiting biases in AI algorithms and data, which is essential to ensure fairness and reliability in safety-critical applications.
- TÜV in Germany has published «Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications» /63/, which focuses on the verification and validation of AI systems. It emphasizes transparency, explainability, and independent third-party assessment to ensure that AI systems work as intended under different operational scenarios.

7 FURTHER WORK

7.1 Joint guidelines for the petroleum industry

Havtil's regulations refer to a wide range of standards and guidelines, both Norwegian and international. Most players choose to follow these, since otherwise they must demonstrate that alternative approaches are just as good or better. The use of common standards and guidelines contributes to a harmonized level of safety in the petroleum industry. However, as described in 2.2 and 5.3, so far no standards or guidelines have been prepared specifically for the safe use of AI in the petroleum industry. To maintain a harmonized level of safety and reduce the burden of proof on the individual player, it would be beneficial if relevant players in the industry could join forces to prepare guidelines that represent best practice for the use of different types of solutions containing AI.

Such an effort should also make it easier to meet the requirements of the EU AI Act. Currently, individual players that want to use AI must specify and respond to the Act's requirements in their own management systems. Such operationalization of high-level requirements is usually labour-intensive, but it should be possible to reduce the workload on each individual organization if the industry makes a joint effort.

7.2 Automated detection of unsafe conditions

A recurring theme in this report is that increased automation can make it difficult for operators to understand when it is necessary to activate barriers manually. The introduction of AI is expected to exacerbate this problem. The challenge associated with human detection of unsafe conditions indicates that the industry should explore the possibilities for automated detection of unsafe conditions caused by AI.

7.3 Qualification and maintenance of AI-based systems

For some AI systems, the results produced will not be deterministic. This means that if you perform several tests with the same input data, you will not necessarily get the same results, which can make it difficult to qualify, validate, and even maintain software that contains AI.

The challenge related to maintenance is that re-running tests to check that you get the same results as before, so-called regression testing, is the most common way to check that there are no unwanted side effects in connection with software modifications.

This may also restrict where AI can be introduced, and the industry should explore how this challenge can be solved.

8 REFERENCES

/1/	Framework HSE Regulations, 18.12.2023, rammeforskriften_e.pdf (havtil.no).
/2/	Management Regulations, 18.12.2023, styringsforskriften_e-1.pdf (havtil.no).
/3/	Activities Regulations, 18.12.2023, aktivitetsforskriften_e-1.pdf (havtil.no).
/4/	Facilities Regulations, 18.12.2023, innretningsforskriften_e-1.pdf (havtil.no).
/5/	Technical and Operational Regulations, 18.12.2023, teknisk_og_operasjonell_forskrift_e2.pdf (havtil.no).
/6/	ICT Security – Robustness in the Petroleum Sector – Regulations and Supervisory Methodology, DNV GL 2019-082, 24.02.2020.
/7/	Principles for barrier management in the petroleum industry – Barrier memorandum, Petroleum Safety Authority Norway 2017.
/8/	Integrated and unified risk management in the petroleum industry, Petroleum Safety Authority Norway, June 2018.
/9/	Oil & Gas UK Guidelines on Qualification of Materials for the Abandonment of Wells, issue 2.
/10/	NOG 070, Guidelines for the Application of IEC 61508 and IEC 61511 in the Norwegian Petroleum Industry (recommended SIL requirements). Guideline, Norwegian Oil and Gas Association, 2020.
/11/	NORSOK-I-002, Industrial automation and control systems, 2021.
/12/	DNV-RP-A203, Technology Qualification.
/13/	DNV-RP-A204, Assurance of digital twins.
/14/	DNV-RP-0317, Assurance of data collection and transmission in sensor systems.
/15/	DNV-RP-0497, Assurance of data quality management.
/16/	DNV-RP-0510, Assurance of data driven applications.
/17/	DNV-RP-0513, Assurance of simulation models.
/18/	DNV-RP-0665, Assurance of machine learning applications.
/19/	DNV-RP-0671, Assurance of AI-enabled systems.
/20/	D. Gupta, M. Shah (2022). A comprehensive study on artificial intelligence in the oil and gas sector. Environ Sci Pollut Res Int. 2022 Jul; 29(34):50984-50997. doi: 10.1007/s11356-021-15379-z (Retracted).
/21/	N. Aissani, B. Beldjilali, D. Trentesaux (2009), Dynamic scheduling of maintenance tasks in the petroleum industry: a reinforcement approach. Eng Appl Artif Intell 22(7):1089–1103. https://doi.org/10.1016/j.engappai.2009.01.014.

/22/	M. Alhashem (2019), Supervised machine learning in predicting multiphase flow regimes in horizontal pipes. Society of Petroleum Engineers - Abu Dhabi International Petroleum Exhibition and Conference 2019. ADIP 2019. https://doi.org/10.2118/197545-ms.
/23/	Bello O, Teodoriu C, Yaqoob T, Oppelt J, Holzmann J, Obiwanne A (2016). Application of artificial intelligence techniques in drilling system design and operations: a state of the art review and future research pathways. Society of Petroleum Engineers - SPE Nigeria. Annual International Conference and Exhibition. https://doi.org/10.2118/184320-ms.
/24/	R. Brelsford (2018). Repsol launches big data, AI project at Tarragona refinery. https://www.ogj.com/refining-processing/refining/operations/article/17296578/repsol-launches-big-data-ai-project-attarragona-refinery.
/25/	Luca Cadei, Marco Montini, Fabio Landi, Francesco Porcelli, Vincenzo Michetti, Matteo Origgi, Marco Tonegutti, Sylvain Duranton (2019). Big data advanced analytics to forecast operational upsets in upstream production system. Society of Petroleum Engineers - Abu Dhabi International Petroleum Exhibition and Conference 2018. ADIPEC 2018. https://doi.org/10.2118/193190-ms.
/26/	S. Chaki, A.K. Verma, A. Routray, W.K. Mohanty, M. Jenamani (2014). Well tops guided prediction of reservoir properties using modular neural network concept: a case study from western onshore, India. J Pet Sci Eng 123:155–163. https://doi.org/10.1016/j.petrol.2014.06.019.
/27/	V.L.C. de Oliveira, A.P.M. Tanajura, H.A. Lepikson (2013). A multi-agent system for oil field management. 11th IFAC Workshop on Intelligent Manufacturing Systems. The International Federation of Automatic Control. 1-6.
/28/	Y. Gidh, A. Purwanto, H. Ibrahim (2012). Artificial neural network drilling parameter optimization system improves ROP by predicting/managing bit wear. Society of Petroleum Engineers – SPE Intelligent Energy International 2012(1):195–207. https://doi.org/10.2118/149801-ms.
/29/	F.A.S. Adesina, A. Abiodun, A. Anthony, F. Olugbenga (2015). Modelling the effect of temperature on environmentally safe oil based drilling mud using artificial neural network algorithm. Petroleum and Coal Journal, Volume 57, Number 1, pp. 60-70.
/30/	A. Ahmed, S. Elkatatny, A. Ali (2021). Fracture pressure prediction using surface drilling parameters by artificial intelligence techniques. J Energy ResourceTechnolgy 143(3):20.
/31/	V. Baskaran, S. Singh, V. Reddy, S. Mohandas (2019). Digital assurance for oil and gas 4.0: role, implementation and case studies. In SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, p 20. https://www.scilit.net/publications/17a663acdcaf179b3a98af9ce989be51.
/32/	S. Choubey, G.P Karmakar (2021). Artificial intelligence techniques and their application in oil and gas industry. Artif. Intell. Rev. 54(5):3665-3683.
/33/	Christopher Jeffery and Andrew Creegan (2020). Adaptive drilling application uses AI to enhance on-bottom drilling performance.
/34/	Yue Wang, Sai Ho Chung (2021), Artificial intelligence in safety-critical systems: a systematic review.
/35/	W. Khalid, I. Soleymani, N.H. Mortensen, K.V. Sigsgaard (2021). AI-based maintenance scheduling for offshore oil and gas platforms. Annual Reliability and Maintainability Symposium (RAMS), p. 1-6.
/36/	L. Kirschbaum, D. Roman, G. Singh, J. Bruns, V. Robu, D. Flynn (2020). AI-driven maintenance support for downhole tools and electronics operated in dynamic drilling environments. IEEE Access, 8, 78683-78701.

/37/	European Commission (2019). Ethics guidelines for trustworthy AI. Technical report. European Commission's High-Level Expert Group on Artificial Intelligence.
/38/	Michael Bortz, Kai Dadhe, Sebastian Engell, Vanessa Gepert, Norbert Kockmann, Ralph Müller-Pfefferkorn, Thorsten Schindler, and Leon Urbas (2023). AI in Process Industries – Current Status and Future Prospects. Chem. Ing. Tech, 95, No. 7, 975-988.
/39/	S.W. Choi, E.B. Lee, J.H. Kim (2021). The engineering machine-learning automation platform (EMAP): a big-data-driven AI tool for contractors' sustainable management solutions for plant projects. Sustainability 13(18):365-384.
/40/	Alexander Inzartsev, Alexander Pavin, Alexander Kleschev, Valeria Gribova (2016). Application of artificial intelligence techniques for fault diagnostics of autonomous underwater vehicles.
/41/	S. Heshmati-alamdari, A. Eqtami, G.C. Karras, D.V. Dimarogonas, K.J. Kyriakopoulos (2020). A self-triggered position based visual serving model predictive control scheme for underwater robotic vehicles. Machines, 8, 33.
/42/	Daniel Mitchell, Jamie Blanche, Sam Harper, Theodore Lim, Ranjeetkumar Gupta, Osama Zaki, Wenshuo Tang, Valentin Robu, Simon Watson, David Flynn (2022). A review: Challenges and opportunities for artificial intelligence and robotics in the offshore wind sector
/43/	Alf Inge Molde (2018). Intelligent progress. Norwegian Continental Shelf 1-2018, p32-36.
/44/	Håvard Devold, Roar Fjellheim (2019). Artificial intelligence in autonomous operation of oil and gas facilities. Society of Petroleum Engineers. SPE-197399-MS.
/45/	Dmitry Koroteev, Zeljko Tekic (2021). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. https://doi.org/10.1016/j.egyai.2020.100041. Energy and AI, Volume 3.
/46/	S. Bernardini, F. Jovan, Z. Jiang, P. Moradi, T. Richardson, R, Sadeghian, S. Sareh, S. Watson, A. Weightman (2020). A multi-robot platform for the autonomous operation and maintenance of offshore wind farms. In Autonomous Agents and Multi-Agent Systems (AAMAS) 2020 International Foundation for Autonomous Agents and Multiagent Systems.
/47/	Håvard Devold, Roar Fjellheim (2019). Artificial Intelligence in Autonomous Operation of Oil and Gas Facilities. Society of Petroleum Engineers, SPE-197399-MS.
/48/	Y.K. Dwivedi, L. Hughes, E. Ismagilova, et al. (2021). Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. International Journal of Information Management.
/49/	Ministry of Local Government and Modernisation (2020). National strategy for artificial intelligence.
/50/	EU Artificial Intelligence Act, COM/2024/1689, http://data.europa.eu/eli/reg/2024/1689/oj.
/51/	NEK ISO/IEC 5338:2023. Information technology – Artificial intelligence – AI system life cycle processes.
/52/	NEK ISO/IEC TR 5469:2024. Artificial intelligence – Functional safety and AI systems.
/53/	NS-EN ISO/IEC 22989:2023. Information technology – Artificial intelligence – Concepts and terminology for artificial intelligence.
/54/	NEK ISO/IEC 23894:2023. Information technology – Artificial intelligence – Guidance on risk management.

/55/	NEK ISO/IEC 42001:2023. Information technology – Artificial intelligence – Management system.
/56/	NEK IEC 61508:2010. Functional safety of electrical/electronic/programmable electronic safety-related systems.
/57/	NEK IEC 61511:2016. Functional safety – Safety instrumented systems for the process industry sector.
/58/	Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu and Ibrahim Habli (2021). Guidance on the assurance of machine learning in autonomous systems (AMLAS). Assuring Autonomy International Programme (AAIP). University of York.
/59/	University of Oxford (2022). capAI; A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act.
/60/	Fraunhofer IAIS (2021). Management system support for trustworthy artificial intelligence; A comparative study.
/61/	NIST (2022). Towards a standard for identifying and managing bias in artificial intelligence.
/62/	Consortium of Quality Assurance (2021). Guidelines for quality assurance of AI-based products and services.
/63/	TÜV (2021). Trusted artificial intelligence towards certification of machine learning applications.
/64/	Dr. Harald Rueß, Prof. Dr. Simon Burton (2022). Fraunhofer IKS. Safe AI—How is this possible?
/65/	J. Smith, L. Brown (2023). Artificial Intelligence in Oil and Gas: Safety and Surveillance Applications. Journal of Petroleum Technology, 2023.
/66/	Kizzy Nkem Elliot, Levi Adawari Damingo (2024), Application of artificial intelligence in the oil and gas industry. International Research Journal of Modernization in Engineering Technology and Science. Volume 6, issue 5.
/67/	Y. Zhang, H. Lee (2022). Real-time safety monitoring in offshore wind farms using AI. Renewable Energy.
/68/	Masoud Masoumi (2023). Machine Learning Solutions for Offshore Wind Farms: A Review of Applications and Impacts. J. Mar. Sci. Eng. 2023, 11(10), 1855; https://doi.org/10.3390/jmse11101855.
/69/	Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery.
/70/	Directive 2014/34/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to equipment and protective systems intended for use in potentially explosive atmospheres.
/71/	Training AI requires more data than we have — generating synthetic data could help solve this challenge (theconversation.com).
/72/	Christoffersen, K., & Woods, D. D. (2002). How to make automated systems team players. In Advances in Human Performance and Cognitive Engineering Research (Vol. 2, pp. 1–12). Emerald Group Publishing Limited. https://doi.org/10.1016/S1479-3601(02)02003-9.
/73/	Littman, M., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi Velez, F., Hadfield, G., Horowitz, M., Isbell, C., Kitano, H., Levy, K., Lyons, T., Mitchell, M., Shah, J., Sloman, S., Vallor, S., & Walsh, T. (2021). Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report. Stanford University. http://ai100.stanford.edu/2021-report.

/74/	National Academies of Sciences, Engineering and Medicine. (2022). Human-AI Teaming: State of the Art and Research Needs. The National Academies Press. https://doi.org/10.17226/26355.
/75/	Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. Human Factors, 59(1), 5–27. https://doi.org/10.1177/0018720816681350.
/76/	Hollnagel, E. (2012). Coping with complexity: Past, present and future. Cognition, Technology & Work, 14(3), 199–205. https://doi.org/10.1007/s10111-011-0202-7.
/77/	Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8.
/78/	Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human Factors, 37(2), 381–394. https://doi.org/10.1518/001872095779064555.
/79/	Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors: The Journal of the Human Factors and Ergonomics Society, 52(3), 381–410. https://doi.org/10.1177/0018720810376055.
/80/	Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(5), 728–739. https://doi.org/10.1177/0018720815581940.
/81/	Mosier, K., & Skitka, L. (1996). Human Decision Makers and Automated Decision Aids: Made for Each Other? In R. Parasuraman & M. Mouloua (Eds.), Automation and human performance: Theory and applications (Vol. 40, pp. 201–220). Lawrence Erlbaum Associates, Inc.
/82/	Finomore, V. S., Shaw, T. H., Warm, J. S., Matthews, G., & Boles, D. B. (2013). Viewing the workload of vigilance through the lenses of the NASA-TLX and the MRQ. Human Factors, 55(6), 1044–1063. https://doi.org/10.1177/0018720813484498.
/83/	Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In Automation and human performance: Theory and applications. (pp. 183–200). Lawrence Erlbaum Associates, Inc.
/84/	Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886.
/85/	Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. Human Factors: The Journal of the Human Factors and Ergonomics Society, 56(3), 476–488. https://doi.org/10.1177/0018720813501549.
/86/	Endsley, M. R. (2023). Ironies of artificial intelligence. Ergonomics, 0(0), 1–13. https://doi.org/10.1080/00140139.2023.2243404.
/87/	Endsley, M. R. (2023). Supporting Human-AI Teams: Transparency, explainability, and situation awareness. Computers in Human Behavior, 140, 107574. https://doi.org/10.1016/j.chb.2022.107574.
/88/	Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

/89/	Bach, T. A., Kristiansen, J. K., Babic, A., & Jacovi, A. (2024). Unpacking Human-AI Interaction in Safety-Critical Industries: A Systematic Literature Review. IEEE Access, 12, 106385–106414. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3437190.
/90/	Kim, J. W., & Jung, W. (2003). A taxonomy of performance influencing factors for human reliability analysis of emergency tasks. Journal of Loss Prevention in the Process Industries, 16(6), 479–495. https://doi.org/10.1016/S0950-4230(03)00075-5.
/91/	Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. Psychological Bulletin, 100(3), 349–363. https://doi.org/10.1037/0033-2909.100.3.349.
/92/	Greenlee, E. T., DeLucia, P. R., & Newton, D. C. (2022). Driver Vigilance Decrement is More Severe During Automated Driving than Manual Driving. Human Factors. https://doi.org/10.1177/00187208221103922.
/93/	Naujoks, F., Purucker, C., Wiedemann, K., & Marberger, C. (2019). Noncritical State Transitions During Conditionally Automated Driving on German Freeways: Effects of Non–Driving Related Tasks on Takeover Time and Takeover Quality. Human Factors, 61(4), 596–613. https://doi.org/10.1177/0018720818824002.
/94/	Veitch, E., Christensen, K., Log, M., Valestrand, E., Hilmo Lundheim, S., Nesse, M., Alsos, O., & Steinert, M. (2022). From captain to button-presser: Operators' perspectives on navigating highly automated ferries. Journal of Physics: Conference Series, 2311, 012028. https://doi.org/10.1088/1742-6596/2311/1/012028.
/95/	Endsley, M. R., Bolté, B., & Jones, D. G. (2003). Designing for situation awareness: An approach to user-centered design. Taylor & Francis.
/96/	Sheridan, T. B. (2021). Human Supervisory Control of Automation. In G. Salvendy & W. Karwowski (Eds.), Handbook of Human Factors and Ergonomics (5th ed., pp. 736–760). Wiley. https://doi.org/10.1002/9781119636113.ch28.
/97/	Wickens, C. D., & Carswell, C. M. (2021). Information processing. In G. Salvendy & W. Karwowski (Eds.), Handbook of Human Factors and Ergonomics (5th ed., p. 1603). John Wiley & Sons, Incorporated.
/98/	ISO. (2019). ISO 9241-210:2019 Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems. International Organization for Standardization.
/99/	Hoem, Å. S., Veitch, E., & Vasstein, K. (2022). Human-centred risk assessment for a land-based control interface for an autonomous vessel. WMU Journal of Maritime Affairs, 21(2), 179–211. https://doi.org/10.1007/s13437-022-00278-y.
/100/	OROK. (2018). S-002:2018+AC:2021 Working environment. Standard Norway.
/101/	van de Merwe, K., Mallam, S., Engelhardtsen, Ø., & Nazir, S. (2023). Towards an approach to define transparency requirements for maritime collision avoidance. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67(1), 483–488. https://doi.org/10.1177/21695067231192862.
/102/	van de Merwe, K., Mallam, S., Nazir, S., & Engelhardtsen, Ø. (2024). The Influence of Agent Transparency and Complexity on Situation Awareness, Mental Workload, and Task Performance. Journal of Cognitive Engineering and Decision Making, 18(2), 156–184. https://doi.org/10.1177/15553434241240553.
/103/	Norman, D. A. (1990). The "problem" with automation: Inappropriate feedback and interaction, not "over-automation." Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 327(1241), 585–593. https://doi.org/10.1098/rstb.1990.0101.

/104/	van Doorn, E., Horváth, I., & Rusák, Z. (2021). Effects of coherent, integrated, and context-dependent adaptable user interfaces on operators' situation awareness, performance, and workload. Cognition, Technology & Work, 23(3), 403–418. https://doi.org/10.1007/s10111-020-00642-z.
/105/	Bach, T. A., Kristiansen, J. K., Babic, A., & Jacovi, A. (2023). Unpacking Human-AI Interaction in Safety-Critical Industries: A Systematic Literature Review (arXiv:2310.03392). arXiv. http://arxiv.org/abs/2310.03392.
/106/	Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation Usage Decisions: Controlling Intent and Appraisal Errors in a Target Detection Task. Human Factors, 49(3), 429–437. https://doi.org/10.1518/001872007X200076.
/107/	Ezenyilimba, A., Wong, M., Hehr, A., Demir, M., Wolff, A., Chiou, E., & Cooke, N. (2023). Impact of Transparency and Explanations on Trust and Situation Awareness in Human–Robot Teams. Journal of Cognitive Engineering and Decision Making, 17(1), 75–93. https://doi.org/10.1177/15553434221136358.
/108/	Oliveira, L., Burns, C., Luton, J., Iyer, S., & Birrell, S. (2020). The influence of system transparency on trust: Evaluating interfaces in a highly automated vehicle. Transportation Research Part F: Traffic Psychology and Behaviour, 72, 280–296. https://doi.org/10.1016/j.trf.2020.06.001.
/109/	Warden, T., Carayon, P., Roth, E. M., Chen, J., Clancey, W. J., Hoffman, R., & Steinberg, M. L. (2019). The National Academies Board on Human System Integration (BOHSI) Panel: Explainable AI, System Transparency, and Human Machine Teaming. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63(1), 631–635. https://doi.org/10.1177/1071181319631100.
/110/	van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. Human Factors, 66(1), 180–208. https://doi.org/10.1177/00187208221077804.
/111/	Strauch, B. (2018). Ironies of Automation: Still Unresolved After All These Years. IEEE Transactions on Human-Machine Systems, 48(5), 419–433. https://doi.org/10.1109/THMS.2017.2732506

APPENDIX A. EXAMPLE: AI-DRIVEN PREDICTIVE MAINTENANCE FOR OFFSHORE DRILLING PLATFORMS

Background

Offshore drilling operations are central to the energy sector but come with complex challenges and inherent risks due to the operational environment, the mechanical complexity of the drilling equipment, and the high costs associated with equipment failures and downtime. In this context, predictive maintenance appears to be a relevant strategy to anticipate and prevent failures before they occur.

AI implementation

Implementing an AI-powered predictive maintenance system on offshore drilling platforms involves deploying a network of sensors and IoT devices to continuously monitor the condition of critical equipment components in real time. This data relates to vibration, temperature, pressure, and other operational parameters that are critical to the drilling equipment's function.

Machine-learning algorithms analyse this data, learning from historical maintenance reports, failure occurrences, and operational anomalies to anticipate potential failures or maintenance needs before a critical failure occurs. These algorithms can not only detect patterns that indicate emerging faults, but also suggest optimal maintenance schedules, thus reducing unnecessary inspections and repairs and minimizing downtime.

Benefits

1. By anticipating equipment failures before they occur, the risk of accidents and hazardous conditions for operators is reduced.
2. Downtime due to unplanned maintenance and repairs is minimized, ensuring that drilling operations run more smoothly and efficiently.
3. AI-powered predictive maintenance can lead to significant cost savings by reducing the need for corrective maintenance and extending the life of critical drilling equipment.

Challenges

- The seamless integration of AI systems with existing offshore infrastructure.
- Ensuring the reliability and accuracy of AI predictions to avoid false positives or negatives. This will be particularly critical if the AI-powered system also determines the maintenance intervals.
- Addressing cybersecurity concerns related to IoT and AI implementation.
- The training of personnel to interact with and respond to AI-driven maintenance recommendations.
- Ensuring that AI implementations comply with existing regulations and standards for safety-critical systems in the offshore industry.

About DNV

We are a global quality-assurance and risk-management company with a presence in over 100 countries. Our purpose is to safeguard lives, assets, and the environment. With our unique technical expertise and independence, we assist our customers in improving safety, efficiency, and sustainability.

Whether we're approving a new ship design, optimizing energy production from a wind farm, analysing sensor data from a gas pipeline, or certifying a food producer's value chain, we help our customers make good and right decisions and increase confidence in their business, products, and services. The world is changing. We can influence developments. Together, we will tackle the global challenges and transitions we face.