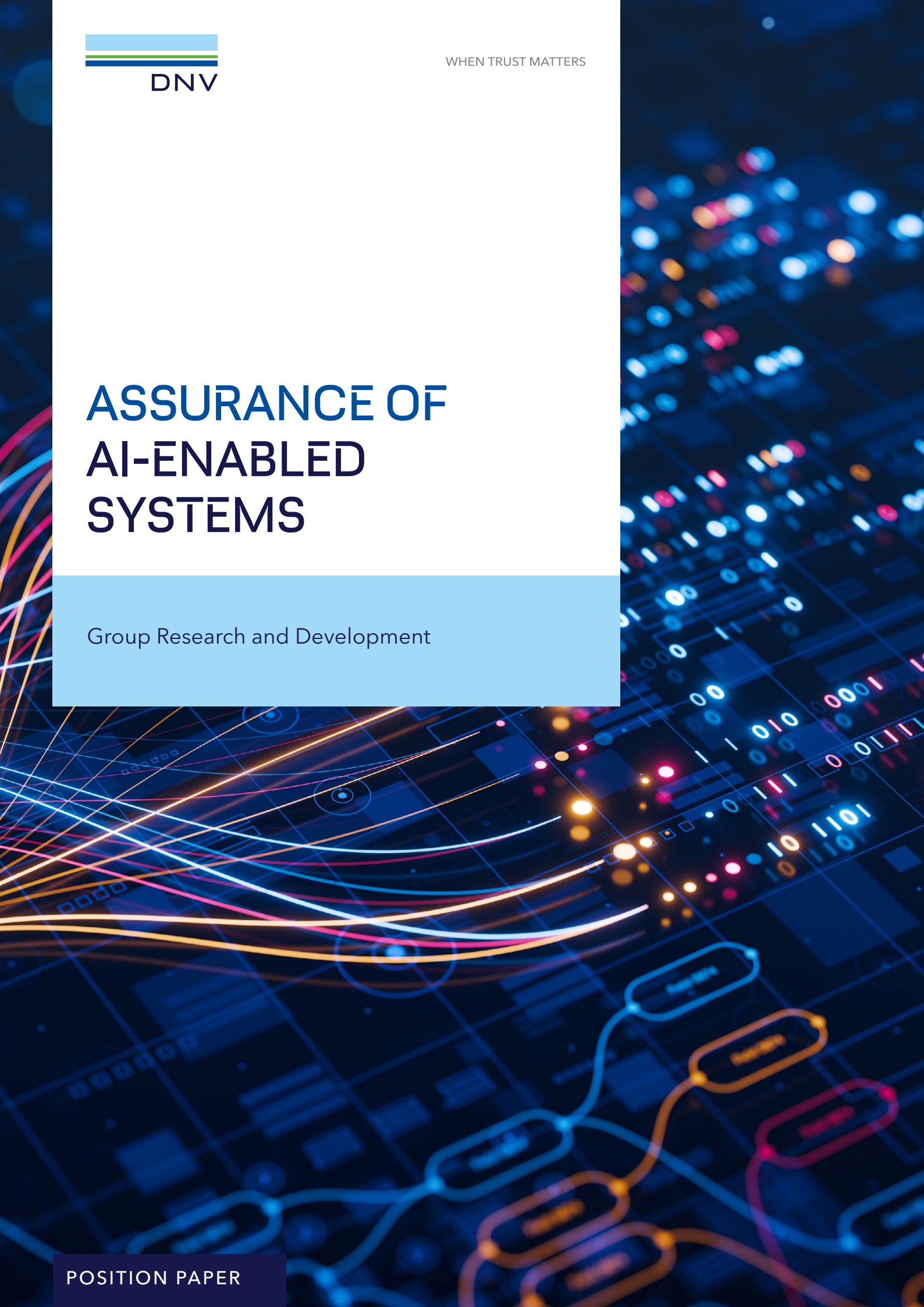
DNV

WHEN TRUST MATTERS

ASSURANCE OF AI-ENABLED SYSTEMS

Group Research and Development



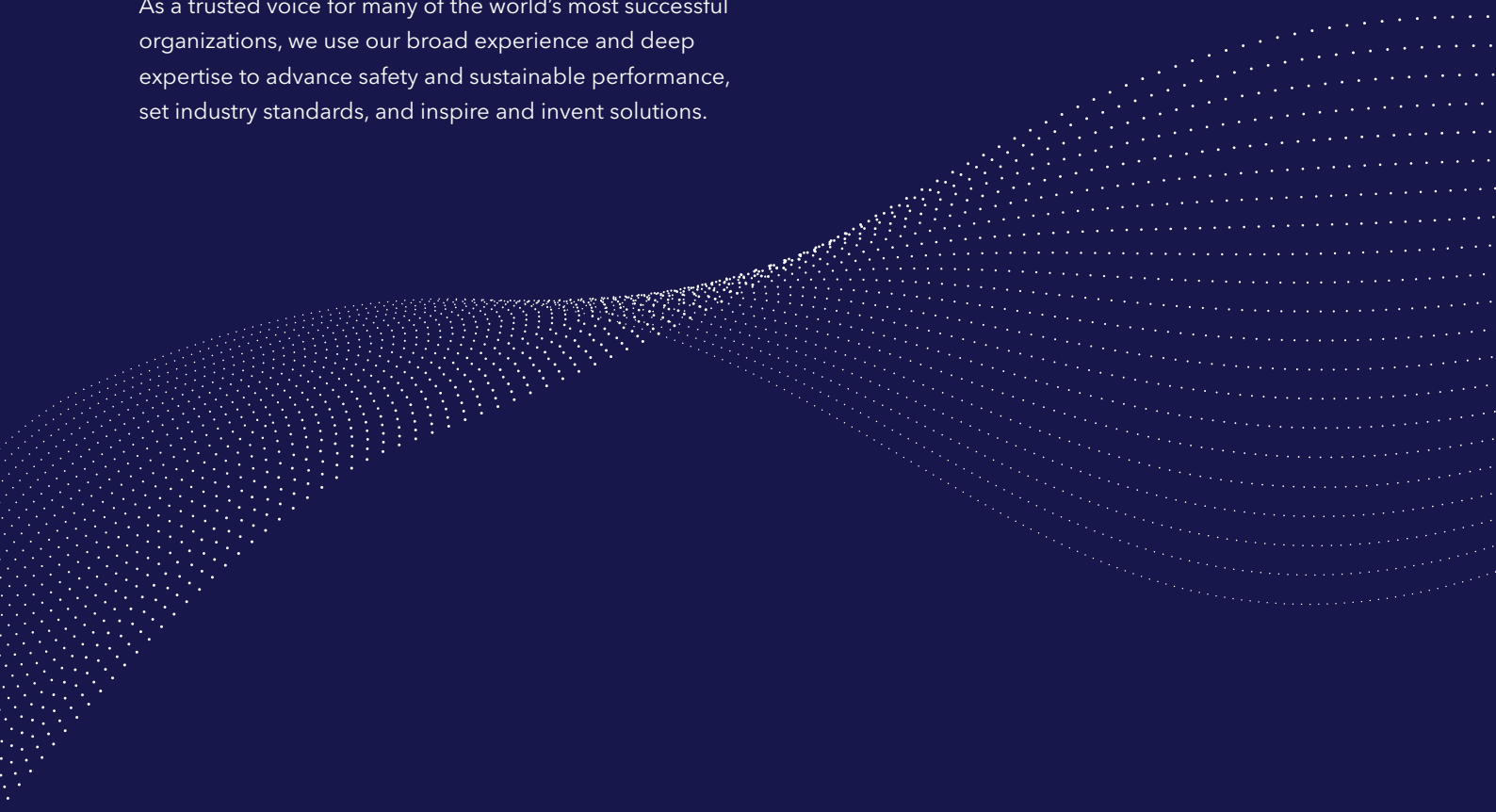
POSITION PAPER

About DNV

DNV is an independent assurance and risk management provider, operating in more than 100 countries, with the purpose of safeguarding life, property, and the environment.

Whether assessing a new ship design, qualifying technology for a floating wind farm, analysing sensor data from a gas pipeline, or certifying a food company's supply chain, DNV enables its customers and their stakeholders to manage technological and regulatory complexity with confidence.

As a trusted voice for many of the world's most successful organizations, we use our broad experience and deep expertise to advance safety and sustainable performance, set industry standards, and inspire and invent solutions.



Layout: ETN Grafisk / Erik Tanche Nilssen AS

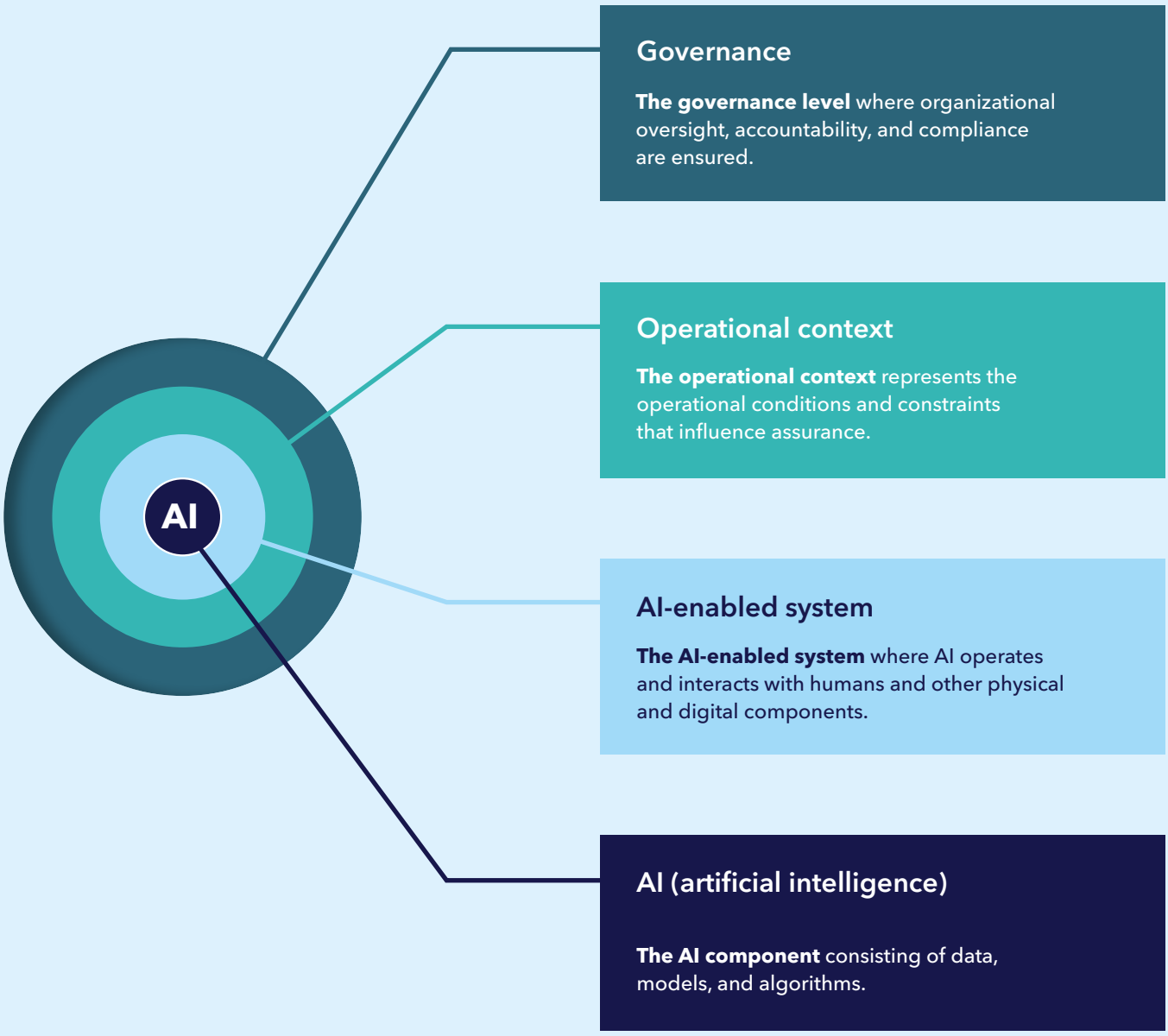
Images: p. 1, 3, 22: Gettyimages, p. 6, 10, 12, 14, 17, 19, 24: Shutterstock

CONTENTS

1 INTRODUCTION TO ASSURANCE OF AI

FIGURE 1

Trustworthy AI means trust in the AI technology, the system where it is used, the operational context, and how it is governed



How can we assure trustworthy AI in high-risk, real-world systems?
AI is being embedded in systems that affect our safety, economy, and daily lives. From self-driving cars to medical diagnosis, AI is affecting decisions with real and critical consequences. To understand if AI is trustworthy and assure its trustworthiness, we must consider the system containing AI: the AI-enabled system.

We define an *AI-enabled system* as a system that contains or relies on one or more AI components /1/. It is where AI operates in the real world by interacting with humans, machines, and digital and physical infrastructures.

When AI fails, it is rarely just a "glitch". It's often the result of hidden risks and biases, unclear responsibilities, or unexpected interactions between AI, people, and the environment where it operates. In addition, these systems evolve through learning, updates, and changing contexts, making traditional assurance methods insufficient.

Imagine a robotaxi that stops instantly when a child chases a ball into traffic. It uses AI to detect the hazard faster than any human could do.

But what if the AI misreads the scene in heavy rain? Or it fails to recognize a cyclist because it was trained on incomplete data? These split-second decisions rely on sensors, algorithms, and connectivity all working perfectly. The same system that saves lives today could kill someone tomorrow.

*AI-enabled systems introduce new and dynamic **risks** that require adequate assurance methods. Establishing trust in AI necessitates (see Figure 1):*

- *trust in the **AI technology** and the infrastructure that supports it,*
- *trust in the **system** where it is used,*
- *trust in the **operational context** of the system, and*
- *trust in the **governance**.*

A robotaxi isn't just a car. It's an AI-enabled system embedded in a complex, dynamic environment. It may behave unpredictably and lead to severe real-world consequences. It therefore requires assurance. Assurance builds confidence across technical, operational, and societal layers, to ensure safety, reliability, and public trust.

> Assurance is the grounds for justified confidence. It is a structured process that provides evidence that something (a product, system, or service) meets the requirements, expectations, or goals of its stakeholders.

In this position paper we present our views on the new risks, challenges, and potential solutions for assurance of AI-enabled systems:

- **AI introduces systemic risks:** These are linked to: (1) weaknesses in the AI itself and (2) the emergent effects that reshape system behaviour, alter human-system interactions, and impact safety and security /2/.
- **AI creates new assurance challenges:** Managing AI risks poses new challenges due to system complexity, frequent updates, distributed responsibilities, and the inherent variability and stochasticity in AI /1/.
- **Proven methods exist:** Assurance methods are available to manage AI risks, by drawing on modern risk science, decades of experience across safety-critical, cyber-physical, and digital systems. This can be matched with operational approaches for AI such as MLOps /3,4/ for effective assurance of AI.

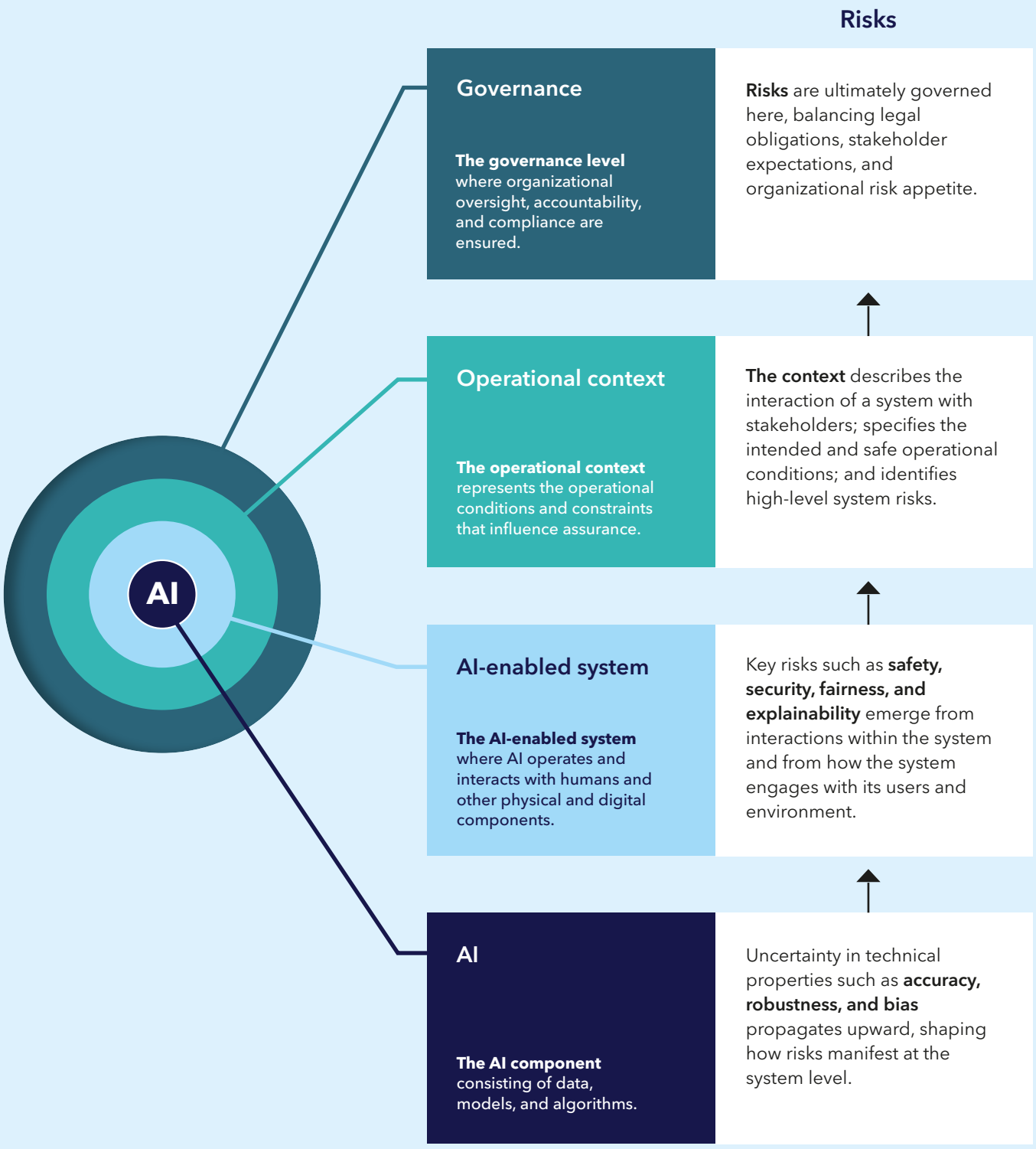
Together, this is pointing towards a new paradigm: **assurance as a continuous, adaptive, system-wide, and evidence-based process**, beyond static checks.

2 AI BRINGS NEW SYSTEM RISKS

When AI is introduced into a system, it fundamentally alters the risk landscape, both changing existing risks and introducing entirely new ones. AI does not operate in isolation. It is being embedded into complex systems such as autonomous vessels, medical devices, and industrial control systems, where its integration transforms the original system risks /2/.

FIGURE 2

Illustration of risks at different levels



While AI can enhance efficiency and safety, it also brings new **AI risk sources** such as incomplete or biased data, poor explainability, frequent updates, low robustness in novel situations, and vulnerabilities to security breaches /5/. These sources do not cause harm directly but influence system risks, such as collisions in maritime operations or misdiagnoses in healthcare, especially when human oversight fails to intervene effectively. To effectively manage risks from these new sources, new assurance approaches also need to be used.

AI risks are **systemic and context-dependent**, emerging from the interactions in the AI-enabled system, with humans and other parts of the system and its environment. For instance, poor AI accuracy might not be a problem in isolation, but when embedded in the system operation combined with inadequate human-machine interactions from poor transparency, it can escalate into critical failures /5/.

Figure 2 illustrates that risks do not only *propagate* from one level to the next. Risks can also *emerge* at each level /6,7/.

- **The AI component:** Some risks are direct AI properties, such as inaccuracy, bias or poor performance in edge cases (robustness) /2,7/.
- **The AI-enabled system:** Risks emerge from how the AI and its properties impact the interaction with other parts of the system, such as sensors, digital and physical components, or human operators during system operation. These emergent risks may in turn lead to outcomes threatening safety, security, or fairness /5,6,8/.
- **The operational context:** High-level system risks are identified in the operational context level through the specific use cases in which AI-enabled systems interact with stakeholders /1/.
- **The governance level**: Finally, risks are managed also at the governance level where legal obligations, stakeholder expectations and organizational risks need to be balanced. But AI also brings challenges to the governance process itself, as new methods, tools, and processes, and hence new competence is needed. The inability to apply an effective governance process is thus also a risk /1,9/.

This means that effective risk management of AI-enabled systems must account for (1) risks that propagate from lower-level components and subsystems, (2) risks that emerge from system-level interactions, (3) risks that emerge from the operational context, and (4) risks related to inadequate governance and system management /10/.

The presence of risk also drives the need for confidence and consequently **assurance**. Assurance provides justified confidence that an AI-enabled system will perform according to stakeholders' objectives (e.g. safety, security, reliability), and in compliance with regulatory and operational expectations, even in the face of uncertainty /11,12/.

3 AI CHALLENGES CONVENTIONAL ASSURANCE AND RISK MANAGEMENT PRACTICES

We presented in Chapter 2 how AI brings new systemic risks. In addition, we argue that the nature of developing, using, and governing AI brings additional challenges to how assurance and risk management is carried out.

Assurance and risk management become challenging with AI-enabled systems that may evolve through learning, exhibit new emergent behaviours, and that depend heavily on data quality and context. This makes it difficult to define complete, static requirements upfront.

Managing the new risks and nature of AI poses challenges at technical, systemic, context, and governance levels (see Figure 3) /1,3,4,13,14/. Here we summarize some of these challenges.

FIGURE 3

AI brings challenges to assurance and risk management (not exhaustive)

Component		System		Context	Governance	
Stochasticity	Uncertainty	Emergent behaviour	Human-AI interaction	Environmental variability	Distributed responsibilities	Compliance
Frequent updates	Data quality	Adaptive behaviour	Integration	Socio-technical constraints	Intellectual property and transparency	Ethical and societal implications

3.1 Challenges at the component level

Challenges at the component level are linked to the AI technology.

(1) Frequent updates and data drift

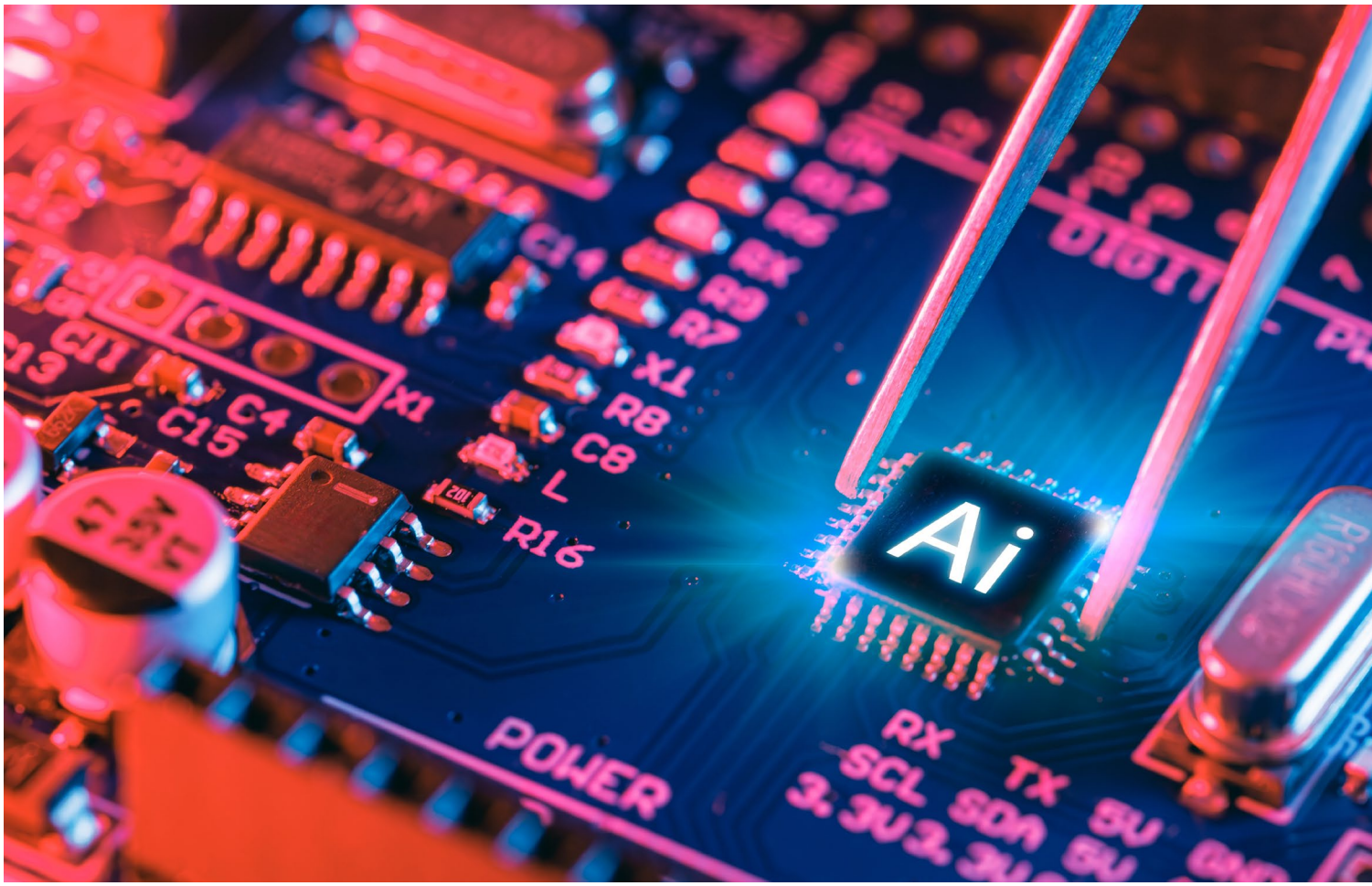
Example: *The robotaxi's perception model is retrained weekly using new urban driving data. A recent update improves pedestrian detection, but also degrades performance in fog, invalidating prior safety claims.*

AI data drift, retraining, and continuous learning lead to frequent updates, which can covertly change the AI risk picture and invalidate static risk assessments and prior safety claims.

(2) Inherent stochasticity and variability

Example: *In identical weather and lighting, the robotaxi sometimes brakes abruptly for a plastic bag and sometimes ignores it, complicating deterministic behaviour.*

AI models are probabilistic by design, sometimes producing variable outputs even for identical inputs. This challenges verification methods that rely on reproducibility and deterministic logic, but more importantly, it challenges over-confident safety claims and narrow safety margins assuming deterministic and accurate system behaviors that are not robust to the inherent variability.



(3) Uncertainty in AI decisions

Example: *The robotaxi struggles to respond consistently to small, fast-moving objects at night, as the video becomes blurry (noisy data leading to aleatory uncertainty). Later, it encounters a road completely blocked by a snowdrift, which is an event never seen during training, and it fails to recognize the hazard (lack of data leading to epistemic uncertainty).*

Decisions made in AI-enabled systems are influenced by hidden statistical relationships in data rather than explicit rules. This creates two types of uncertainty that must be explicitly managed:

- **Epistemic uncertainty**: due to incomplete knowledge (e.g., scenarios missing from training data, or causal relationships represented by probabilistic distributions rather than deterministic rules)
- **Aleatory uncertainty**: due to noise or inherent randomness (e.g., noisy training data)

Regardless of the type of uncertainty, it may lead to unwanted variability or errors that impact system behaviour. A transparent treatment of both types of uncertainty is critical for credible risk claims and stakeholder trust /4/.

(4) AI model performance depends on dynamic data properties

Example: *Pedestrian detection was trained on data where some cyclists were mislabelled as pedestrians. The model learns incorrect patterns and struggles to distinguish between the two, leading to unsafe braking or path planning when encountering cyclists.*

Model performance depends on three key properties of training and validation data:

- **Quality**: accuracy, completeness, and integrity of data (e.g., correct labels, no corruption)
- **Representativeness**: coverage of relevant populations, operating conditions, and edge cases (e.g., diverse weather, lighting, road types)
- **Temporal validity**: alignment with the current and evolving deployment context (e.g., new urban layouts, vehicle types)

The assurance challenge is that each dimension is **dynamic**:

- Quality can degrade through **changes in how the data is captured**
- Representativeness can erode as **use context shifts**
- Temporal validity is threatened by **data drift and changing environments**

3.2 Challenges at the AI-enabled system level

Challenges at the systemic level are linked to the AI-enabled *system*.

(1) Added complexity from non-linear, adaptive and emergent system behaviors.

Example: *When multiple robotaxis coordinate at an intersection, they create unexpected traffic oscillations. This is emergent behaviour not seen in individual testing.*

AI introduces complexity not just from its algorithms, but from how it interacts with other system components, humans, and its operational environment. AI-enabled systems can exhibit **emergent behaviours**: behaviours that emerge during operation at a system level (here: in traffic) and not from an individual part (here: a single robotaxi). These emergent behaviours can include unintended capabilities or failures that arise only during operation due to interactions between components, humans, and data /5,15,16/.

These behaviours cannot be fully predicted during design. Hazards can emerge from system level interactions (see /6/ for discussion).

Assurance must therefore adopt a **systems approach**, which captures how AI interacts across multiple system levels /9/. AI's adaptive nature means system behaviour can evolve after deployment, requiring dynamic assurance methods that can capture emerging risks over time /4/.

(2) Integration and operational challenges

Example: *The robotaxi relies on cloud-based HD maps; and a network outage disables localization.*

Integrating AI into industrial systems introduces operational challenges beyond technical compatibility. AI components often depend on specialized infrastructure, such as cloud platforms or embedded hardware. This adds complexity and new failure modes. Additionally, assurance must address the need for operational monitoring of the AI component itself and how the AI interacts with humans and other parts of the system /17/.

(3) Dynamic objectives and evolving system behaviors

Example*: The robotaxi adapts to changing training objectives and additional training data, prioritizing ride comfort over speed and agility, but this also changes braking patterns, which then invalidates the original safety case built on conservative response times.*

AI-enabled systems operate in environments where objectives and operational contexts may evolve. As the AI learns and adapts to new data and conditions, its behaviour may change to a degree that invalidates elements of the original assurance case. Assurance must therefore be continuous and lifecycle-oriented, with mechanisms for monitoring, re-evaluation, and the updating of assurance and safety cases, as well as operational constraints /4/.

(4) Human-AI interaction as a safety-critical concern.

Example: *The safety driver fails to intervene in time.*

Humans often act as fallbacks or overseers of AI-enabled systems in high-risk environments /18,19,20/. However, as AI-enabled systems gain autonomy, the assumption that humans can always intervene effectively becomes increasingly flawed. Human roles need to be clearly defined, supported, and aligned with system capabilities, even in specific operational scenarios. This includes:

- ensuring that operators understand AI outputs,
- operators have the authority to override decisions when necessary,
- operators are not burdened with unrealistic expectations of vigilance or fault recovery,
- and finally, the system must be designed for this.

3.3 Challenges at the context level

Challenges at the context level are linked to operational conditions, stakeholder interactions, and external constraints that influence assurance.

(1) Environmental variability and dynamic operational conditions

Example: *The deployment of a robotaxi trained in cities with wide roads is challenging in a dense European city, and requires retraining to restore safety.*

Operational environments are rarely static. Variability in physical conditions (e.g., weather, lighting, terrain) and digital conditions (e.g., network connectivity) can significantly alter system behaviour. This challenges assurance because risk assessments based on fixed assumptions may not hold under changing conditions, leading to safety and reliability gaps /1/.

(2) Stakeholder expectations and socio-technical constraints

Example: *When a robotaxi is deployed, only the safety of the passengers and pedestrians are considered but not the effect of introducing an autonomous vehicle on the flow of the traffic.*

AI systems operate within socio-technical ecosystems where human roles, organizational processes, and cultural norms shape risk. Divergent stakeholder expectations can create governance gaps and operational friction, increasing the likelihood of unsafe decisions or accountability failures. Trustworthiness depends on mapping stakeholder roles, understanding human–machine interdependencies, and ensuring transparent communication and impact assessment /1,15,29/.

3.4 Challenges at the governance level

Challenges at the governance level are related to the governance and management of the AI-enabled systems.

(1) Distributed responsibilities challenges

Example: *When the robotaxi hits a pothole it didn't detect, blame is disputed: did the sensor provider or operator fail? No single party owns end-to-end safety.*

The key challenge of distributed responsibilities in AI-enabled system is defining and enforcing who is accountable for what in a multi-stakeholder and decentralized system when AI failures or harms occur /1,20/.

- **Multiple stakeholders, unclear boundaries:** AI-enabled systems involve developers, data providers, integrators, deployers, and end users. In distributed architectures (e.g., federated learning, multi-agent systems), it's hard to see where one party's responsibility ends and another's begins.
- **Unfair attribution of blame to human operators**: When AI-enabled systems operate autonomously, it is unfair and misleading to hold human operators (e.g., safety drivers) accountable for failures they cannot reasonably prevent. If the system acts without clear human override capability or transparent alerts, blaming the human becomes a moral and systemic failure. This undermines trust and discourages meaningful human oversight.
- **Complex technical chains**: As AI becomes more complex, tracing the root cause (or several root causes) of a problem is difficult. A decision error could be due to a faulty sensor in edge AI, a parameter update in the central server, or a bug in the local model of a node. Determining who should take responsibility for corrective actions or compensations is challenging.
- **Incentive – responsibility mismatch**: In some cases, a mismatch may appear between responsibilities of the actors (e.g., developers and deployers). The parties with more control or benefit may not bear corresponding responsibilities, while those with limited influence are overly burdened.

(2) Compliance challenges

Example: *The robotaxi company decides to enter a new geographical market in which existing regulation (e.g. EU AI act) classifies it as high-risk, which requires new documentation, real-time monitoring, and third-party audits not previously implemented.*

The core compliance challenge of AI-enabled systems is aligning AI-enabled systems with diverse and evolving standards and regulations.

- **Diverse and evolving rules, standards and regulations**: AI laws and regulations (e.g., EU AI Act) vary by jurisdiction, industry, and how they handle different AI risk levels. In addition, rules and standards (e.g., CEN/CENELEC Joint Technical Committee 21) need to keep up with the rapid advancements of AI. Navigating these different requirements creates compliance complexity.
- **Cross-border compliance risks**: AI-enabled systems often operate globally, but regulations have territorial scope. The EU AI Act's /21/ extraterritorial application (applying to non-EU companies targeting EU users) means organizations must comply with multiple jurisdictions' rules leading to overlapping obligations.
- **Technical feasibility gaps**: Many compliance demands (e.g., traceability of algorithmic decisions, bias mitigation, data privacy) require technical solutions that are not always mature. For instance, proving explainability for complex models or auditing distributed AI-enabled systems is technically challenging and resource-intensive /18/.
- **Documentation and accountability**: Regulations like the EU AI Act /21/ require extensive documentation (e.g., risk assessments, technical records, stakeholder consultations) for high-risk AI. Maintaining and demonstrating this documentation, especially across distributed teams or supply chains, is a heavy administrative load.

(3) IP and transparency challenges

Example: *The robotaxi's AI uses proprietary models and training data whose origins are unclear.*

The key IP and transparency challenges in AI governance are about balancing innovation protection with transparency/explainability and accountability.

- **IP and accountability challenges** in AI arise from unclear data rights, conflicting incentives between model protection and transparency/explainability, and ambiguous liability in collaborative systems. This creates legal and ethical risks across the AI lifecycle. In addition, incomplete disclosure of data origin and preprocessing, along with poor traceability in distributed systems hinders auditability and accountability.

- **Technical opacity**: AI "black box" models are hard to explain and current explainability methods are often insufficient for effective oversight.

(4) Ethical and societal Implications

Example: *The robotaxi detects an ambiguous object in the road, possibly a small child or animal, forcing an instant choice between swerving (endangering passengers inside the robotaxi) or proceeding (risking harm to the object in the road).*

One of the core challenges of AI-enabled systems is balancing AI's innovation potential with the need to mitigate harm to individuals, groups, and society, while addressing vague ethical standards and conflicting values.

- **Ambiguous ethical standards**: Ethical principles (fairness, privacy, autonomy, non-maleficence) lack universal, actionable definitions. For example, "fairness" could mean equal treatment across groups or equitable outcomes, leading to conflicting priorities in AI design (e.g., a hiring AI balancing gender parity vs. merit-based selection).

- **Changes to human autonomy and agency**: Over-reliance on AI (e.g., automated decision-making in healthcare, education, or law) risks reducing human control. Governance must strike a balance: leveraging AI efficiency without undermining human judgment or stripping individuals of the right to challenge AI decisions /18/.

- **Conflicting global values** (e.g., privacy vs. security, innovation vs. labour rights) and **rapid AI adoption** can lead to societal harm, including job loss, democratic erosion, and social fragmentation.

4 THE FOUNDATIONS FOR ACHIEVING TRUSTWORTHY AI

In the previous chapter, we presented the risks and challenges that make achieving safe and trustworthy AI difficult. However, while AI-enabled systems pose unique risks requiring novel assurance solutions, there is an existing foundation of safety and trustworthiness principles that can be adapted and extended.

The solution is integrated and adaptive assurance that moves beyond static compliance to a lifecycle-wide, evidence-based approach. This can be achieved by combining methods from risk and assurance science with digital practices such as DevOps/MLOps, and industry best practices from existing safety-critical and high-risk operations.

While there are methods developed for safety-critical domains, their practical application to AI requires structured adaptation. Furthermore, many new methods and frameworks reside in academic or highly specialized literature, making access hard. To help bridge this gap we highlight some of the key elements that make up the foundations for trustworthy AI.

4.1 Modern risk and assurance approaches

Assurance and risk are intrinsically linked. Assurance provides confidence, which can be needed on key inputs in a risk assessment. And conversely, a risk assessment can provide evidence needed into assurance. Decades of development in safety-critical industries like maritime, energy, nuclear, and aerospace have advanced this discipline to handle digitalization, autonomy, and the complexity introduced by AI.

In our experience, these modern risk and assurance approaches can be incorporated through three interconnected elements:

- A system model, built using systems theory, to represent the AI-enabled system and its operational context.
- A risk model, applying uncertainty-based assessment and modular risk principles to break down complex systems with their complex and emergent risks into manageable parts across system levels /3,7/.
- An assurance case (see Figure 4), which serves as the formal argumentation and knowledge model that links claims (e.g., "the system is safe") to relevant evidence through structured reasoning, assumptions, and traceable justification /3,6,8,10,15,22/.

These are the core elements needed for effective assurance of AI-enabled systems.

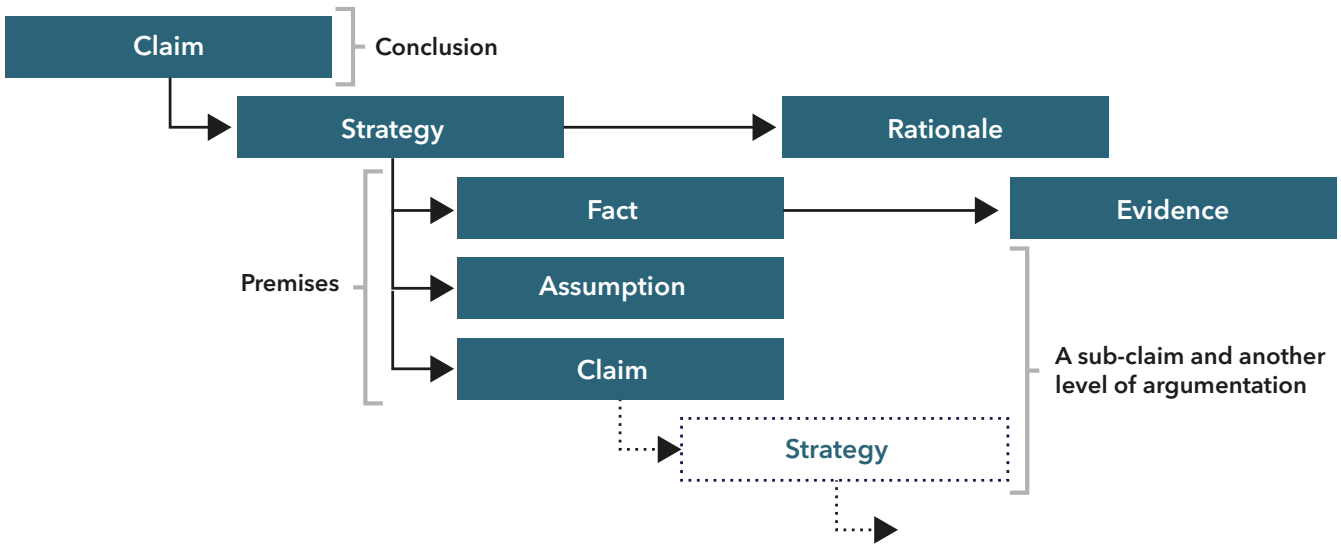
Furthermore, by using a modular approach, risks can be assessed and assured at different levels, from component to system, without requiring a single, monolithic argument. This supports distributed responsibility, frequent updates, and evolving use contexts, while enabling reuse of proven modules /23/.

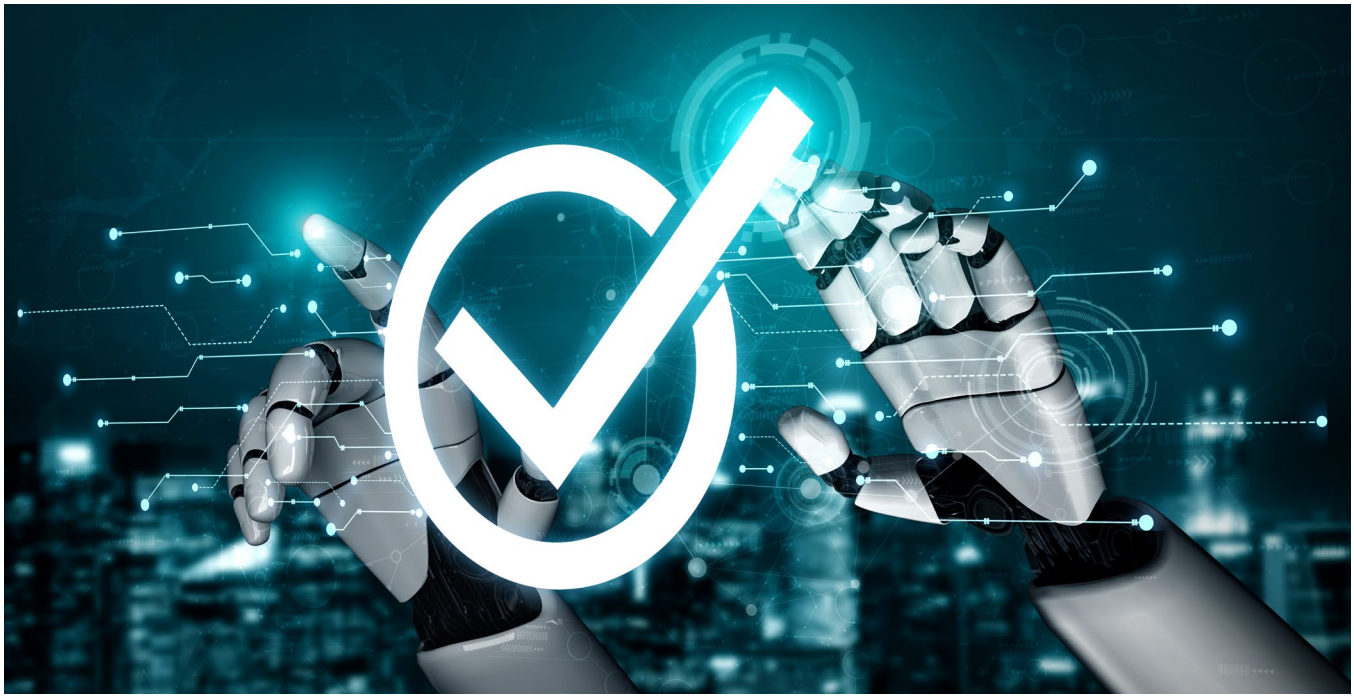
There is extensive experience applying risk and assurance approaches across industries such as automotive, maritime, energy, and healthcare, where different risk paradigms exist but also common principles emerge. Right now, assurance frameworks from safety-critical domains are being adapted to address autonomy and adaptive AI behaviour.

For example, the Safety 4.0 book /24/ provides a framework for safety in novel cyber-physical systems. It overcomes limitations of traditional approaches when applied to software and AI, integrating systems engineering, technology qualification (TQ), and risk-informed assurance cases. It uses systems-theoretic methods (e.g., STPA, FAST), uncertainty propagation, and evidence-based argumentation. From this foundation, DNV-RP-0671 Assurance of AI-enabled systems /1/ was developed to tailor the approach specifically to AI.

FIGURE 4

The Assurance case. This is a way to formalize the justification of confidence, structuring how system understanding, risks, and evidence combine into a coherent, auditable argument.





4.2 The assurance process

For any method, tool or framework to be useful, it must be incorporated into a practical process. The assurance process described in "DNV RP-0671" /1/ starts with defining stakeholder-informed claims, such as "the system is safe" or "the system is fair", that reflect intended behaviour and trustworthiness. It proceeds by identifying risks that could undermine these claims, followed by deriving risk-informed system-level requirements. These are decomposed into technical specifications for AI components, covering data, models, and algorithms. Evidence is then generated through testing, monitoring, and analysis, and structured into assurance cases that link claims to evidence via logical argumentation. The process emphasizes traceability, continuous validation across the lifecycle, and regulatory alignment, in order to enable justified confidence in the AI-enabled system's performance and safety.

An important part of assuring AI-enabled systems is to derive requirements and follow up that they are met. Requirements may be on *processes* (what should be done and how), or on the *properties of the system, sub-systems or components* (how something should function) /1/. It is the level of risk that shapes how requirements are formulated and what is sufficient for meeting requirements (acceptance criteria). Figure 5 illustrates the AI risks and requirements at four distinct levels: the governance level, the operational context, the AI-enabled system, and the AI component.

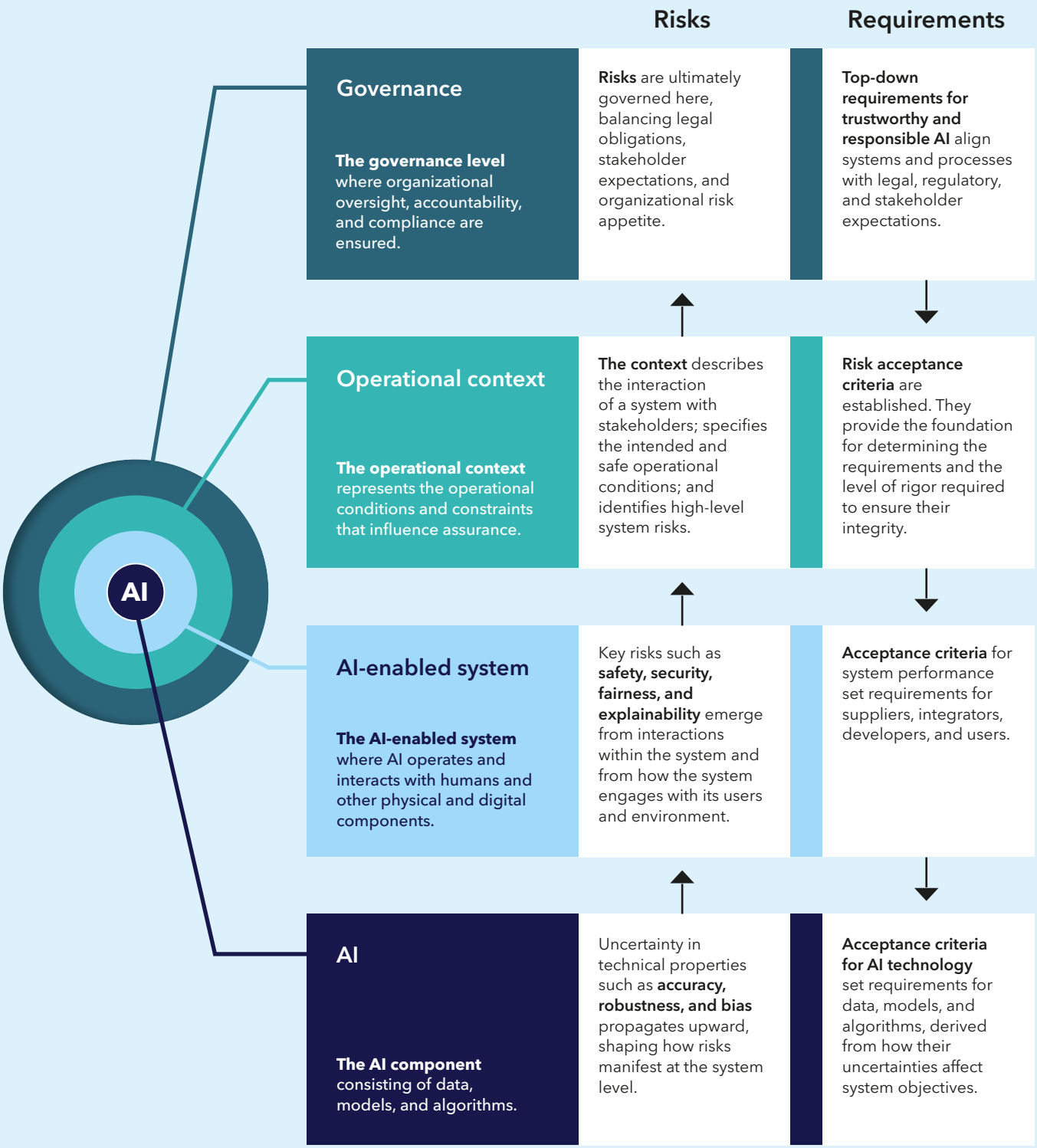
- Governance of AI-enabled systems is concerned with ensuring that legal and business requirements are met. "ISO/IEC 42001:2023 - AI management systems" /25/ is an example of a standard that sets requirements to this process.

- An example of a system-level claim can be that "the system is safe", or "the system is fair". These claims must be transformed to formal requirements, first on the system, and then to specific sub-systems or components. "IEC 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems" /26/ is an example of a standard that sets requirements to system-level safety.

- Based on the operational context and system-level requirements, the AI component must achieve certain properties. For example, the model must have an average accuracy of 99%. "ISO/IEC TS 4213:2022 Information technology – Artificial intelligence – Assessment of machine learning classification performance" /27/ is a technical specification that provides requirements for how the performance of a certain type of AI component (ML classifiers) should be measured. It does not set *acceptance criteria* for these measured properties, that will depend on how the AI is used within the system. For example, if the level of autonomy is increased, then requirements may need to be updated.

FIGURE 5

Illustration of risks and requirements at different levels



4.3 Enabling practices

The following practices are especially relevant for assurance of AI:

- **Technology qualification**
  Technology Qualification (TQ) is a systematic process to demonstrate that a new technology performs reliably in its intended environment, ensuring safety, performance, and compliance /28/. It involves hazard identification, testing, and validation across lifecycle stages. For AI, TQ can support risk management by verifying robustness, explainability, and resilience to bias or drift. This strengthens AI assurance by providing structured evidence for trustworthiness, and enabling justified confidence in AI decisions under uncertainty /1,29/. In particular, TQ would be useful to qualify the barriers and guardrails used to safeguard AI.

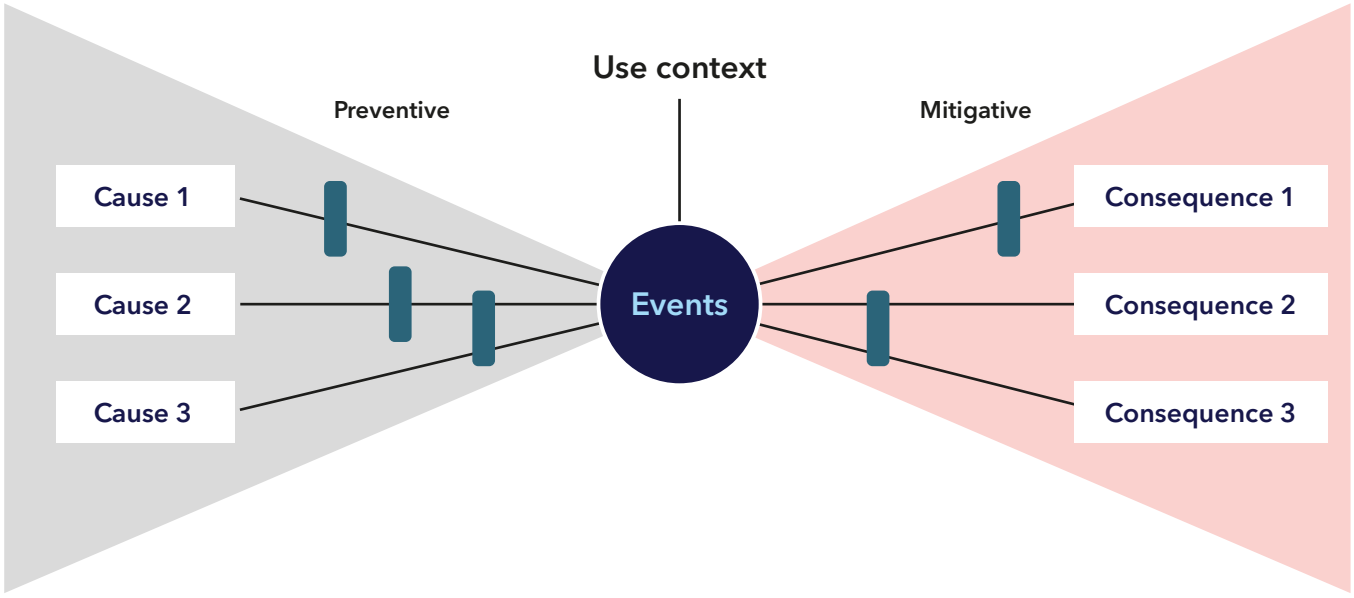
- **DevOps/MLOps**
  DevOps or MLOps practices enable continuous integration, testing, and real-time performance monitoring throughout the AI lifecycle. These practices support adaptive assurance by detecting drift, ensuring robustness, and maintaining trust throughout operation /4/.

- **Emerging responsible AI governance frameworks**
  Responsible AI governance has a focus on stakeholder involvement in the early stages of implementation to ensure their needs and interests are met. This early involvement is important for harm prevention, fairness, accountability, explainability, AI literacy, privacy, and human-AI calibration /20,30/.

- **Barrier management**
  The bowtie model is commonly used in risk management together with the barrier philosophy (see Figure 6). It illustrates how initial mechanisms or causes may lead to an unwanted event, which in turn may result in various consequences. Barriers are introduced as control mechanisms to either prevent the unwanted event from occurring (preventive barriers) or to mitigate its consequences should it occur (mitigative barriers). Effective risk management requires understanding the relationships and uncertainties associated with the causes, the unwanted event, the resulting consequences, and the performance and reliability of barriers put in place.

FIGURE 6

A bow tie with preventive and mitigative barriers



- **Causes and Events**: Capturing the relevant causes and events is important. Taking a systems perspective with stakeholder input (e.g., through STPA) helps to identify context-dependent initiating events in AI-enabled systems.

- **Consequences**: Assessing severity and likelihood is challenging with novel technology. It requires system and risk models, as well as a way to quantify uncertainty when the strength of knowledge varies. In cases where event likelihoods cannot be estimated reliably, the analysis must account for this uncertainty. The uncertainty-based risk perspective /31,32/ provides the method for this.

- **Barriers**: New technological barriers and barrier elements are available for AI. For example, out-of-distribution detection and LLM guardrails. The effectiveness of these barriers may change during operation and should be monitored.

How do we know that the relevant events, causes and consequences have been identified? And how can we be confident that the risks we have captured and the way we manage them actually lead to the system properties we aim for, such as safety? Providing this confidence is at the heart of assurance.

4.4 The result: continuous, context-aware, evidence-based assurance

Together, modern methods and practices can provide a robust foundation for managing AI risks. DNV's ambition to ensure trustworthy AI builds upon this foundation. Based on the COMPASS project /33/ and DNV-RP-0671 (Recommended Practice for Assurance of AI) /1/, we provide an approach that moves beyond static compliance to assurance of AI-enabled systems that provides justified confidence and trustworthiness across technical, operational, and governance layers. Assurance needs to remain valid at all times. It is determined by the frequency of updates and being relevant to the risks, and we denote this as ***continuous assurance***. Such an approach is:

- **Continuous,** being dynamic and adapting to frequent updates, data drift, and evolving operational conditions through real-time monitoring and feedback loops (e.g., MLOps).
- **Context-aware**, recognizing that risks emerge from interactions between AI, humans, and environments.
- **Evidence-based**, using structured assurance cases to link claims (e.g., "the system is safe") to verifiable data from testing, monitoring, and uncertainty quantification.

This gives a view into the future of assuring AI-enabled systems, not as a one-time check, but continuous justified confidence.

5 CONCLUSION AND CALL TO ACTION

To ensure trustworthy AI in high-risk, real-world systems, organizations must adopt **adaptive, system-wide, and lifecycle-integrated assurance practices** to unlock the full potential of AI in high-risk applications.

- **Recognize AI risks as systemic**: Risks do not come from the AI component alone, but also emerge from its integration into the system and interactions between AI, humans, other components, and the surrounding environment.
- **Acknowledge the new challenges**: Complexity, dynamism, distributed responsibility, and uncertainty demand a paradigm shift in risk and assurance thinking.
- **Adopt appropriate methods that are ready for AI**: Leverage assurance cases, continuous monitoring, and uncertainty-aware system and risk modelling.
- **Comply with evolving regulatory requirements:** Start early and make the governance structure ready for AI to ease compliance.
- **Address ethical and societal concerns:** All risks with AI are not yet uncovered, understood, and captured by regulation. For responsible use of AI, it is important to take ethical and societal concerns seriously. Failing to do so may also cause significant reputational risk.

Regulatory frameworks like the **EU AI Act** /21/ mandate such approaches for high-risk AI-enabled systems. Proactive adoption of these methods is a **strategic enabler of innovation, safety, and stakeholder trust**.

We recommend organizations to start now by integrating assurance into the AI lifecycle, build cross-functional governance, and invest in the tools and competencies needed for trustworthy AI.

REFERENCES

/1/ DNV, "*DNV-RP-0671: Assurance of AI-enabled systems*," DNV Recommended practice, Norway, 2023.

/2/ "*Painting the AI Risk Picture*," DNV, Blog Post, 2023.

/3/ A. Hafver, I. Jakopanec, S. Eldevik, D. V. Lindberg, and F. B. Pedersen, "*Enabling confidence: Addressing uncertainty in risk assessments*," DNV GL Strategic research & innovation position paper, DNV GL, Høvik, Norway, 2016.

/4/ "*AssuranceOps: Building Trust in Evolving AI*," DNV, Blog Post, 2026, forthcoming.

/5/ O. I. Haugen, "*A Systems Approach to Modelling Emergent Behaviour of Maritime Control Systems Using the Composition, Environment, Structure, and Mechanisms (CESM) Metamodel*," Group Research and Development, DNV, Høvik, Norway, 2024.

/6/ O. I. Haugen, A. Karlsen, S. Mearns Cargill, and J. van Tiggelen, "*Beyond Component Failures: Safety Challenges in Complex Maritime Control Systems*," DNV, Høvik, Norway, 2025.

/7/ A. Hafver, D. A. Kuruge and A. Vats, "*Four hurdles of industrial AI – and paths to overcome them,*" DNV Technology Insights, DNV, Høvik, Norway, 2026.

/8/ O. I. Haugen, "*Safety Assurance of Complex Systems, Part 1-3*," Group Technology and Research White Paper, DNV GL, Høvik, Norway, 2019.

/9/ A. Hafver, F. B. Pedersen, I. Jakopanec, L. Oliveira, J. Domingues, S. Eldevik, and D. V. Lindberg, "*Maintaining Confidence: Dynamic Risk Management for Enhanced Safety*," Group Technology and Research Position Paper, DNV GL, Høvik, Norway, 2017.

/10/ A. Hafver, "*What is the meaning and origin of risk in complex intelligent systems – and why does it matter?*," 67th ESREDA seminar proceedings, 25-26 September 2025, Høvik, Norway.

/11/ A. Hafver, C. Ferreira, C. Agrell, D. McGeorge, E. A. Hektor, F. B. Pedersen, M. van der Meulen, O. I. Haugen, S. Eldevik, and T. Myhrvold, "*On the Meaning of Assurance*," Group Technology and Research, DNV, Høvik, Norway, 2021.

/12/ O. I. Haugen, "*Building Confidence: An Ontological Approach to Assurance of Safety-Critical Systems*," Group Research and Development, DNV, Høvik, Norway, 2025.

/13/ A. Hafver, F.B. Pedersen, "*Beyond Words? The Possibilities, Limitations, and Risks of Large Language Models*," DNV, Blog Post, 2024.

/14/ A. Hafver, L. Zhao, W.-s. Bao, F. Wu, S. El Mekkaoui, A. Babic, E. Y. Liu, "*Safe, responsible and effective use of LLMs*," DNV Technology Insights, DNV, Høvik, Norway, 2024.

/15/ A. L. St. Clair, Ø. Smogeli, A. Ødegårdstuen, J. A. Glomsrud, S. Eldevik, and C. Nadeau, "Trustworthy Industrial AI Systems," Group Technology & Research Position Paper, DNV GL, Høvik, Norway, 2019.

/16/ J. H. Holland, "*Complexity: A very Short Introduction*," Oxford University Press, 2014.

/17/ T. A. Bach, A. Babic, N. Park, T. Sporsem, R. Ulfsnes, H. Smith-Meyer, and T. Skeie, "*Using LLM-Generated Draft Replies to Support Human Experts in Responding to Stakeholder Inquiries in Maritime Industry: A Real-World Case Study of Industrial AI*," DNV, Høvik, Norway, 2024.

/18/ J. A. Glomsrud, A. Ødegårdstuen, A. L. St. Clair, and Ø. Smogeli, "Trustworthy versus Explainable AI in Autonomous Vessels," in *Proc. International Seminar on Safety and Security of Autonomous Vessels*, Helsinki, Finland, Sep. 2019, pp. 1–10.

/19/ T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective," *Int. J. Hum.–Comput. Interact.*, vol. 40, no. 5, pp. 1123–1145, 2024.

/20/ T. A. Bach, "*Implementing Trustworthy and Responsible AI for Critical Infrastructure: The Role of Stakeholders in Fostering and Maintaining Trust*," DNV, Blog Post, 2025.

/21/ European Parliament and Council of the European Union, "*Regulation (EU) 2024/1689 on Artificial Intelligence (the Artificial Intelligence Act)*," Official Journal of the European Union, vol. L, no. 168, pp. 1–140, 2024.

/22/ O. I. Haugen, "*Integrating assurance and risk management of complex systems*," Group Research and Development, DNV, Høvik, Norway, 2024.

/23/ D. McGeorge and J. A. Glomsrud, "A modular risk concept for complex systems," in *Proc. 44th Int. Conf. Comput. Safety, Reliability Security (SAFECOMP) Position Papers*, Stockholm, Sweden, Sep. 2025.

/24/ M. van der Meulen and T. Myhrvold, Eds., "*Demonstrating Safety of Software-Dependent Systems: With Examples from Subsea Electric Technology,*" DNV, Høvik, Norway, 2022.

/25/ *ISO/IEC 42001:2023, Information technology — Artificial intelligence — Artificial intelligence management system — Requirements*, 2023.

/26/ *IEC 61508:2010, Functional Safety of Electrical/Electronic/ Programmable Electronic Safety-Related Systems (E/E/PE or E/E/PES)*, 2nd ed., 2010.

/27/ *ISO/IEC TS 4213:2022, Information Technology — Artificial Intelligence — Assessment of Machine Learning Classification Performance*, 2022.

/28/ DNV, "*RP-A203 Technology qualification*," DNV Recommended practice, Norway, 2019/2021.

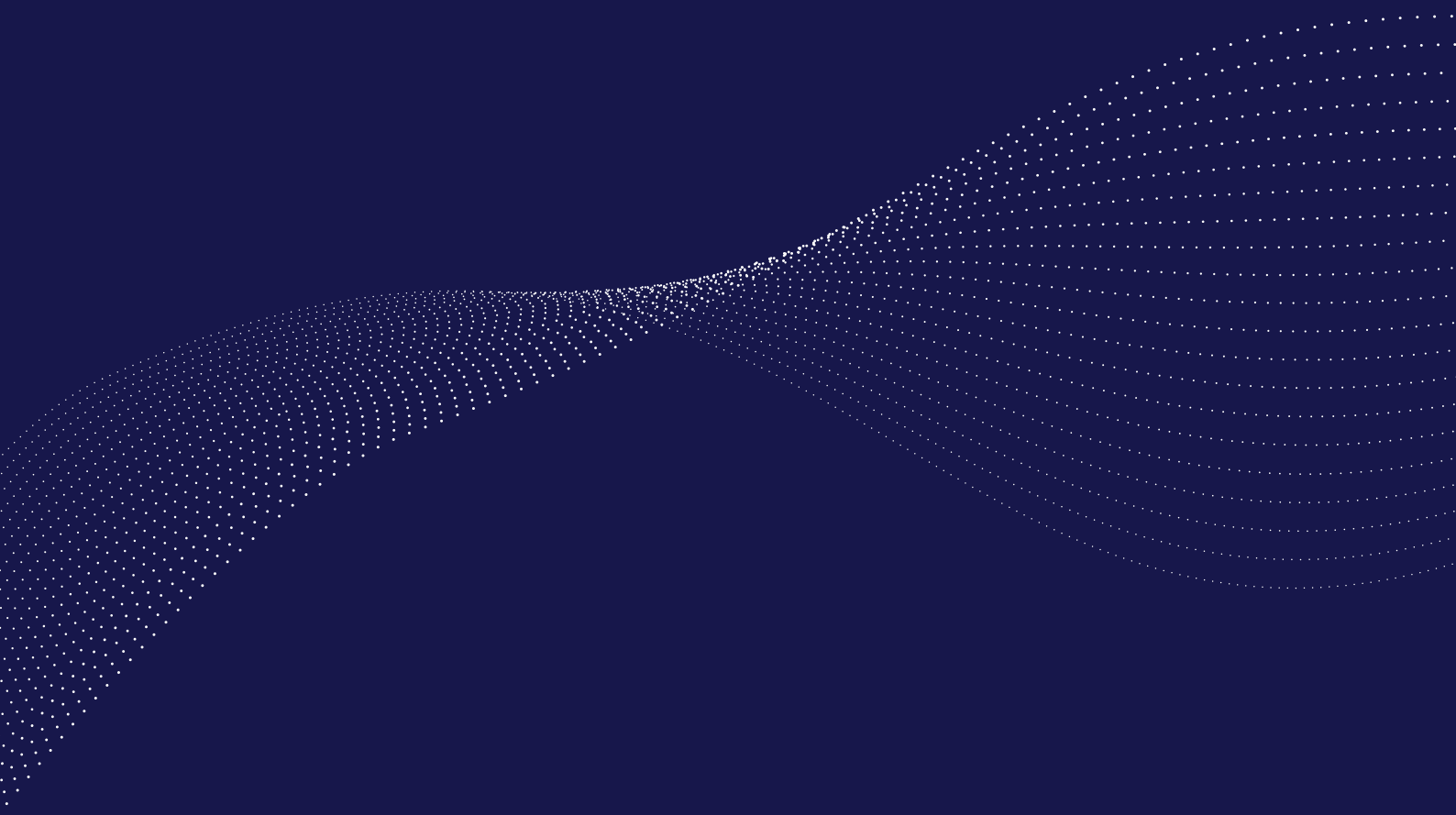
/29/ A. L. St. Clair, D. McGeorge, J. A. Glomsrud and A. Hafver, "*Assurance in the digital age*," Group Research and Development Position Paper, DNV, Høvik, Norway, 2022.

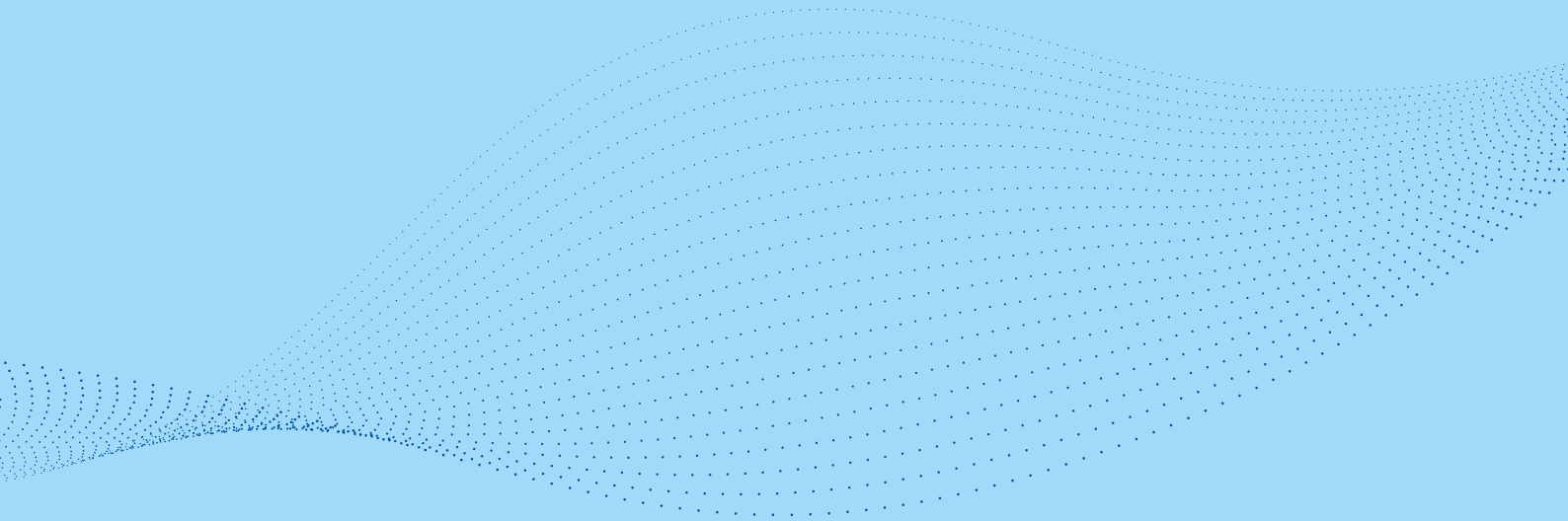
/30/ T. A. Bach, M. Kaarstad, E. Solberg *et al.*, "Insights into suggested Responsible AI (RAI) practices in real-world settings: a systematic literature review," *AI Ethics*, vol. 5, pp. 3185–3232, 2025.

/31/ *ISO 31000:2018, Risk management — guidelines,* 2018/2023.

/32/ T. Aven, "*The Science of Risk Analysis. Foundation and Practice*," Taylor & Francis, 2020.

/33/ COntinuous, Modular and adaPtive Assurance of complex SystemS (COMPASS) project. Project number 355828. Innovasjonsprosjekt i Næringslivet (IPN) project funded by Norwegian Research Council. 2025-2028.





Headquarters:
DNV AS
NO-1322 Høvik, Norway
Tel: +47 67 57 99 00
www.dnv.com