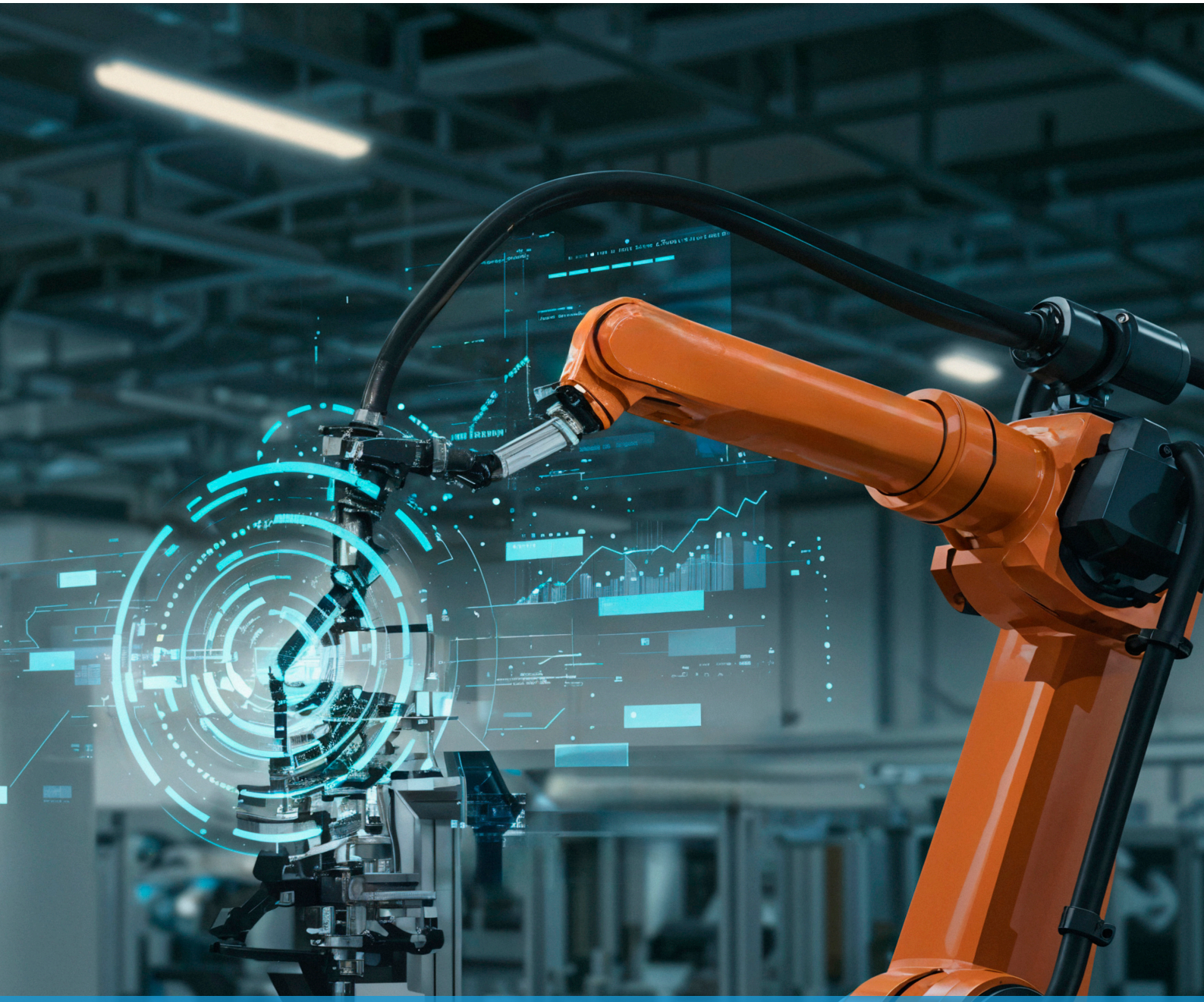
UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION
Progress by innovation

AIM GLOBAL | Global Alliance on AI for Industry & Manufacturing



AI for Industry: A Governance Framework

A Governance Framework for Responsible Deployment, System Level Reliability, Safety, and Controllability

White Paper, July 2026

上海人工智能研究院
Shanghai Artificial Intelligence Research Institute

bsi

I-AIIG
清华大学人工智能国际治理研究院
INSTITUTE FOR AI INTERNATIONAL GOVERNANCE
TSINGHUA UNIVERSITY

Potential Implementing Partner
AI GOVERNANCE FOR HUMANITY LAB

Acknowledgements

This white paper contributes to UNIDO's work on Industrial AI Governance and was prepared with the overall guidance of the UNIDO Directorate of Technical Cooperation and Sustainable Industrial Development (TCS).

Under the leadership of Jason Slater, Chief of the Division for Digital Transformation and Artificial Intelligence (DAI), UNIDO coordinated the preparation of the paper, which was developed as a contribution to discussions on the future of Industrial AI Governance in the context of the World Artificial Intelligence Conference (WAIC) 2026.

The lead author of this white paper was Dr. Xueqi ZHAO, with the main contributors being Dr. Shtefi MLADENOVSKA, Dr. Zilong HE, and Kejda BIDULI. The infographics were designed by Alwin PAUL JOHN.

UNIDO gratefully acknowledges the guidance and support of Ciyong Zou, Deputy to the Director General and Managing Director of the Directorate of Technical Cooperation and Sustainable Industrial Development, and Professor Lan Xue, Dean of the Institute for AI International Governance, Tsinghua University, whose leadership and support contributed to the development and launch of this White Paper.

UNIDO also wishes to acknowledge the valuable contributions of its partners to the preparation and dissemination of this White Paper. We extend our sincere appreciation to the Institute for AI International Governance of Tsinghua University for its collaboration in the development and joint launch of the White Paper; to the British Standards Institution (BSI) for its valuable support as a strategic partner; and to the AI Governance for Humanity Lab for its support as a potential implementation partner in the preparation, launch, and future advancement of this initiative; and to the Shanghai Artificial Intelligence Research Institute (SAIRI) for its contribution to coordinating the launch of the White Paper.

www.unido.org

UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION

White Paper

AI FOR INDUSTRY: A GOVERNANCE FRAMEWORK

A Governance Framework for Responsible Deployment, System Level Reliability, Safety, and Controllability

Vienna, Austria
July 2026

UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION

Table of Contents

Acronyms and Abbreviations

AI	Artificial Intelligence
AIDIN	AI and Digital for Industry Navigator
ALARP	As Low As Reasonably Practicable
APIs	Application Programming Interfaces
APRL	AI Performance Readiness Levels
BSI	British Standards Institution
CCGF	Controllability-Centred Governance Framework
CoE	Centre of Excellence
EU	European Union
GPAI	Global Partnership on Artificial Intelligence
ISID	Inclusive and Sustainable Industrial Development
ISO	International Organization for Standardization
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
ITU	International Telecommunication Union
JTC 1/SC 42	Joint Technical Committee 1, Subcommittee 42 (Artificial Intelligence)
OECD	Organisation for Economic Co-operation and Development
OPERI	Operations Excellence Readiness Index
QA	Quality Assurance
R&D	Research and Development
SC	Subcommittee
SDG	Sustainable Development Goal
SDG 9	Sustainable Development Goal 9: Industry, Innovation and Infrastructure
SMEs	Small and Medium-sized Enterprises
TCS	Directorate of Technical Cooperation and Sustainable Industrial Development (UNIDO)
TRL	Technology Readiness Level
TRLs	Technology Readiness Levels
UN	United Nations
UNIDO	United Nations Industrial Development Organization
WAIC	World Artificial Intelligence Conference

At a Glance

Industrial artificial intelligence (AI) is increasingly used to coordinate production, logistics and safety functions. In these applications however, failures can often times affect supply chains, workers and surrounding communities beyond the deploying firm. Because governance remains fragmented across jurisdictions and industrial sectors, this paper proposes a Controllability-Centred Governance Framework (CCGF) based on six principles: safety, reliability, transparency, accountability, inclusivity, and sustainability. The framework is operationalised through a three-level governance architecture spanning the global, national, and industry/enterprise levels, a three-phase implementation roadmap, and three complementary families of governance instruments: registration and disclosure, tiered risk classification, and enabling public goods.

As the United Nations specialised agency with a unique mandate to promote inclusive and sustainable industrial development, reflected in Sustainable Development Goal 9 (SDG 9), UNIDO could, subject to Member State guidance, available resources, and coordination within the wider United Nations system, support stakeholder convening, technical cooperation activities, and the provision of the global public goods envisaged under this governance architecture, as further specified in Section 3.3. This contribution would build on the functions described in Section 4.2.2. These include UNIDO's mandate for inclusive and sustainable industrial development, its partnerships with international standards bodies and national counterparts, and its role in supporting multilateral cooperation, knowledge exchange, and consensus building.

Key Messages

- As industrial AI is increasingly deployed in infrastructure-critical contexts, failures have the potential to cascade across supply chains and communities. This elevates governance considerations beyond conventional quality-assurance practices alone.
- Governing industrial AI effectively requires a shift from a mindset of competition to one of shared responsibility, which treats safety, reliability, and controllability as shared public goods rather than private compliance costs.
- The Controllability-Centred Governance Framework (CCGF) gives this logic operational substance, translating six principles, namely safety, reliability, transparency, accountability, inclusivity, and sustainability, into a common reference for responsible deployment.
- Given that industrial AI risk crosses organisational and national boundaries, the framework adopts a three-level nested architecture covering the global, national, and industry and enterprise levels. Each level reinforces the others through mutual recognition, information sharing, and coordination.
- A phased five-year pathway proposes an actionable transition pathway, sequencing national registries, conformity assessment with mutual recognition, regulatory sandboxes, sector codes, and, in the longer term, potential for strengthened international cooperation, including possible framework arrangements for industrial AI governance.
- Under its mandate of promoting inclusive and sustainable industrial development, UNIDO could support, facilitate, and contribute to the international coordination, technical cooperation, and the provision of public goods, including testing platforms and relevant databases, in line with Member State requests, available resources, and coordination within the United Nations system.

Foreword

Artificial intelligence is rapidly transforming the way industries produce, innovate, and compete. Across manufacturing value chains, AI is unlocking new opportunities to improve productivity, optimize resource use, strengthen resilience, and accelerate industrial innovation. At the same time, its increasing autonomy and complexity introduce new governance challenges that demand practical, trusted, and internationally aligned approaches to deployment.

Ensuring that industrial AI remains safe, reliable, and beneficial requires more than technological innovation alone. It calls for governance frameworks that enable countries, enterprises, and innovators to adopt AI responsibly while preserving flexibility for continued innovation. This is particularly important for the Global South, where strengthening governance capacities is essential to ensuring that AI becomes a driver of industrial transformation rather than a source of new digital divides.

Developed through extensive collaboration under the AIM Global Alliance on Artificial Intelligence for Industry and Manufacturing, the Governance Framework provides practical guidance for policymakers, regulators, industry leaders, developers, and manufacturers. Rather than prescribing rigid rules, it promotes governance throughout the AI lifecycle by embedding controllability, accountability, transparency, and continuous oversight into industrial AI systems.

The Framework supports countries in strengthening governance capabilities, fostering trustworthy AI ecosystems, and translating governance principles into practical industrial implementation. Together with UNIDO's Industrial AI Action Plan, it provides a pathway for scaling responsible AI adoption across manufacturing sectors. Through UNIDO's global network of AIM Centres of Excellence and technical cooperation programmes, these principles can be adapted into scalable and replicable solutions that respond to diverse national contexts while maintaining shared international values.

I invite governments, industry leaders, researchers, innovators, and development partners to engage actively through AIM Global as we move from dialogue to implementation and from principles to practical solutions.

We must ensure businesses and innovators leverage AI to be makers and not takers of emerging technologies. UNIDO remains committed to working with governments, industry, academia, and international partners to transform governance into practical action, helping ensure that industrial AI is not only innovative, but also safe, accountable, and inclusive for all.

Ciyong Zou
Deputy to the Director General
Managing Director, Directorate of Technical Cooperation
and Sustainable Industrial Development
United Nations Industrial Development Organization

Foreword

For those of us working in science and technology policy, industrial AI presents a familiar challenge but now it plays out in the physical world. Technology routinely outruns the institutions charged with governing it. Yet when AI is embedded in machines, production lines, and supply chains, the stakes change. An error does not remain a data glitch; it propagates through sensors, equipment, operating routines, and the judgement of the workforce, all nested within arrangements that assign authority and responsibility. Industrial AI, in short, must be governed not as a software product but as a sociotechnical system.

In my view, governance should not promise certainty where none yet exists. Its proper role is more modest but more vital: to enable institutions to act under uncertainty, learn from outcomes, and revise their approaches accordingly. This capacity to adapt is itself a form of institutional support for innovation. Governance reduces risk, certainly, but it also shapes the social arrangements and market conditions through which technology is adopted. Neither government nor industry can foresee every contingency. They must, therefore, share knowledge and adjust policy as evidence accumulates.

What I find particularly valuable in this White Paper is its treatment of controllability as a property of the entire sociotechnical system not of the algorithm alone. Controllability depends on observable system behavior, traceable decisions, equipment that responds as intended, and operators who can monitor and, when necessary, override. The proposed Controllability-Centered Governance Framework gives practical expression to this insight through six interconnected principles: safety, reliability, transparency, accountability, inclusivity, and sustainability. Together, they extend responsibility beyond the code to the full industrial setting and its lifecycle.

Governance must also be organized to learn across levels. The report proposes a nested three-level architecture spanning the global, national, and enterprise spheres. Its instruments should be judged pragmatically, by evidence from use. Registration and disclosure can improve visibility; tiered risk classification can focus attention where consequences are greatest; public goods can build shared capability; and sandboxes, alongside continuous monitoring, support iterative adjustment. The proposed AI Performance Readiness Levels (APRL) scale is offered as an analytical tool—one that grades the demonstrated operational reliability of a given AI system within a declared operating domain. It remains a proposal to be tested, calibrated by sector, and validated against operational evidence. It is neither a standard nor a formal UNIDO methodology. Common principles need not yield identical institutions. Countries can coordinate on fundamental safety while pursuing implementation paths suited to their legal frameworks and capacities. Across connected value chains, safety, reliability, and controllability become shared public goods.

Governance capacity, however, remains uneven and capacity building should therefore be understood as integral to global safety. Subject to Member State guidance, available resources, and coordination within the wider United Nations system, UNIDO can help bridge international standards and industrial practice through technical cooperation.

Ultimately, the value of this White Paper will be determined not by the elegance of its framework, but by how its proposals perform in practice and by how experience informs their revision.

Lan Xue
Dean, Schwarzman College, and
Director, Institute for AI International Governance
Tsinghua University

Abstract

Artificial intelligence (AI) is reshaping global manufacturing, moving from task automation to system-level coordination, real-time prediction, and autonomous decision-making across the production process. As it does so, industrial AI is increasingly seen as critical infrastructure, and the risks it generates have become systemic, interconnected, and transnational: an algorithmic error can cascade through tightly coupled production networks and along global value chains before human inspection can intervene. Existing governance, however, remains anchored in competitive development and is structurally insufficiently suited to manage risk of this character. This paper responds by reframing industrial AI governance around shared responsibility and by proposing a Controllability-Centred Governance Framework (CCGF) organised around six core principles: safety, reliability, transparency, accountability, inclusivity, and sustainability. Building on these principles, it develops a three-level nested governance architecture spanning the global, national, and industry and enterprise levels, and it operationalises that architecture through a phased implementation strategy and three concrete instrument families: registration and disclosure, tiered risk classification, and enabling public goods. UNIDO's industrial-development mandate, its neutrality and its country presence together would enable the Organization to support and facilitate the international coordination platform and to contribute to the provision of the public goods that responsible industrial AI requires, in line with Member State requests.

Keywords: *industrial AI governance; controllability; critical infrastructure; shared responsibility; inclusive and sustainable industrial development*

1 Introduction

Artificial intelligence (AI) is rapidly empowering industrial development. Across manufacturing, AI applications are delivering measurable gains in productivity, product quality, and operational resilience, and they are doing so through an expanding repertoire of uses that includes predictive maintenance, process optimisation, autonomous quality control, and AI-driven supply-chain management. AI presents structural challenges for the Global South, particularly where gaps in digital infrastructure, digital skills, access to finance, and innovation ecosystems exist (Labrunie, Fan and Leal-Ayala, 2026). More significantly, the risks that accompany industrial AI have outrun the existing governance toolkit.

Understanding the governance challenge requires recognising three key features of these risks.

- First, they are systemic, since errors propagate rapidly through tightly coupled production networks.
- Second, they are interconnected, since AI increasingly coordinates process control, logistics, energy management, and workforce oversight at once, which expands the risk areas.
- And finally, they are transnational, since an incident originating in one jurisdiction can cascade along global value chains into many others.

Taken together, these characteristics place industrial AI squarely within the domain of critical-infrastructure governance, which demands coordinated international action, shared responsibility for risk management, and institutional arrangements operating at multiple levels.

The prevailing model of AI-for-industrial development is competitive by design. National strategies prioritise first-mover advantage, research and development investment, and the cultivation of nationally competitive ecosystems, and this competitive impulse has undeniably driven innovation. As the foundation for governance, however, this approach has proven incomplete. The same dynamic has produced a fragmented landscape of regulatory gaps and inconsistent standards, with limited cross-border cooperation on safety and reliability, as well as unclear liability in the event of system failure. The institutional response to failure therefore tends to be reactive rather than anticipatory, addressing harm after it has already spread.

This paper argues for a shift from competition to shared responsibility. Industrial AI binds producers, workers, and surrounding communities into a community of shared exposure to common hazards. This idea draws from the disaster-risk governance literature, where joint exposure to a hazard creates a collective obligation to manage it through cooperative institutions. That obligation redirects attention from the race to deploy towards the shared duty to make deployment safe, reliable, and controllable for every party affected by its consequences, including those who do not control the systems themselves. The paper develops this argument in three steps. It first analyses the risks that industrial AI now generates and the reasons conventional quality assurance cannot contain them. It then sets out the governance principles, consolidated in the Controllability-Centred Governance Framework (CCGF). Building on these principles, it develops a systematic three-level governance framework, which spans the global, national, and industry and enterprise levels. Finally, it explores the implementation pathways and governance instruments through which the framework can be put into practice.

2 Governance Rationale and Principles

2.1 From Efficiency Tool to Critical Infrastructure

The description of industrial AI as an efficiency tool no longer captures how these systems function. Contemporary industrial AI performs system-level coordination, real-time predictive analytics, and consequential decision-making with minimal human oversight. In doing so, it substitutes for human judgement in contexts where error carries significant safety, quality, and economic consequences. The relevant policy question is therefore how to govern a technology now embedded across industrial operations.

Industrial AI is becoming increasingly central within the production networks it coordinates and increasingly embedded in complex interdependencies across industrial systems. This means that its failures generate costs that fall well beyond the deploying organisation. When an AI process controller at a petrochemical facility fails, or an AI logistics optimiser routes shipment through an unstable corridor, for instance, the harm spreads across supply chains and communities, and the deploying firm cannot fully internalise it. Risks of this kind are not just operational risks, but matters of public interest, and market mechanisms alone will not govern them adequately. On each of these counts, industrial AI now functions as critical infrastructure rather than a discretionary efficiency tool, which is what makes its governance both necessary and urgent.

The urgency becomes clearer when examining how industrial AI fails. **Three questions are relevant: how AI risks spread, why they are difficult to detect, and why they will grow more complex over time.** Each question isolates a distinct risk mode that traditional quality assurance (QA) – designed for discrete and observable defects – may be unable to manage. Figure 1 consolidates these modes into a single risk framework, and Table 1 sets out, for each one, its underlying mechanism, an illustrative manifestation, and the reason conventional QA fails to detect it.

To answer the first question – **how do AI risks spread?** – they cascade. Modern production networks are tightly coupled, so an algorithmic error at one stage propagates into the next before it can be addressed, and defective output passes through successive quality checks undetected. Because these errors move through physical infrastructure at machine speed, they can outrun the inspection cycles meant to contain them. This is the cascading mode: by the time a defect becomes visible, it has typically moved past the point where inspection could have stopped it.

Addressing the second question – **why these risks are difficult to detect** – the primary challenge is opacity. Some industrial AI models rely on black-box decision logic, meaning that their decision-making process is not observable. Therefore, gradual drift or degradation can accumulate without visible warning until a physical consequence emerges. Mechanical failures usually announce themselves through heat, vibration, or noise, whereas a degrading AI system often gives no comparable signal, especially in continuous operations. A defect-detection model may therefore

lose accuracy while continuing to report normal performance, and a robotic assembler may gradually drift out of tolerance long before the deviation is noticed. This is the stealth mode of risk, and it defeats inspection regimes.

And finally, **why will these risks become more complex over time?** The answer lies in the increasing density of AI deployment. As enterprises extend AI across process control, predictive maintenance, energy management, and logistics - and as these systems are supplied by different vendors with divergent design assumptions - the number and complexity of their interactions increase. These interactions can amplify risk non-linearly and produce behaviours that no single system exhibits on its own. Because risk assessment is typically conducted at the level of individual systems, it cannot capture these compound effects, and the resulting risk surface widens with every additional system introduced. This is the compound mode of risk, and it intensifies as industrial AI deployment deepens.

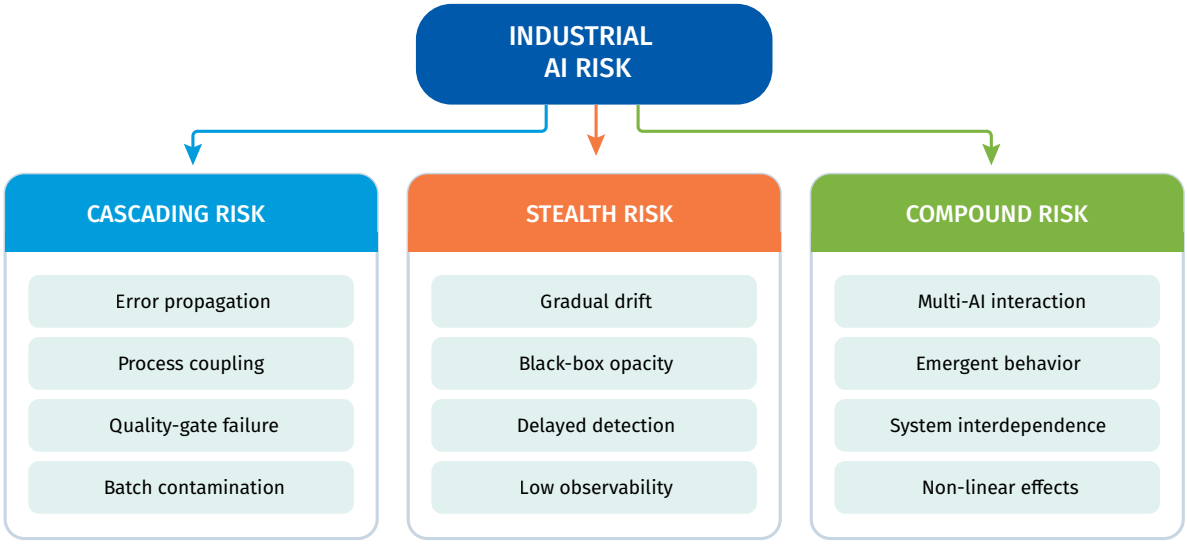


Figure 1. Industrial AI Risk Framework

These modes are not abstract possibilities. They follow from the way modern production systems are built and connected. Plausible examples include process errors in semiconductor fabrication traced to inadequately validated AI models, or supply-chain misrouting by optimisers that fail to anticipate geopolitical disruption. Recognising risks of this character, the EU AI Act classifies certain AI systems used in safety-critical industrial applications and critical infrastructure as high-risk and establishes requirements for conformity assessment, registration where applicable, and post-market monitoring (European Parliament and Council of the European Union, 2024). Transnational governance gaps call for further actions and competitive AI development modalities require a transformation.

Risk Mode	Mechanism	Illustrative Manifestation	Why Conventional QA Fails
Cascading	Tightly coupled production networks transmit algorithmic errors through sequential stages.	An AI process-control error propagates downstream, and defective output passes through successive quality checks.	Consequences cascade through physical infrastructure before human inspection can intervene.
Stealth	Black-box decision logic masks gradual drift or degradation until physical consequences emerge.	A robotic assembler drifts gradually out of tolerance, or a defect-detection model degrades without triggering alerts.	Mechanical failures usually generate warning signals, whereas AI drift often does not, especially in continuous processes.
Compound	Multiple AI systems interact across processes, maintenance, energy, and logistics functions.	Interaction effects between AI components from different vendors amplify risks non-linearly.	Single-system risk assessment cannot capture emerging risks, and the risk surface widens.

Table 1. Three Risk Modes That Warrant Critical-Infrastructure Classification

2.2 Core Principles: The Controllability-Centred Framework

The three risk modes set out in Section 2.1 outlined what any adequate governance regime must be able to do. This section now explores six principles arising from the risk modes just described.

Cascading risk

- where an algorithmic error propagates before human inspection can intervene
- calls for systems to carry explicit safety objectives and remain subject to human override, as well as that responsibility be clearly assigned in case of harms crossing firm boundaries.

Stealth risk

- where performance drifts or degrades silently because of non-observable decision logic
- calls for reliability and transparency respectively.

Compound risk

- where hazards emerge from the interaction of systems supplied by different vendors and operated across different functions
- highlights the need for transparency and clear accountability across the whole sociotechnical system rather than a single component.

Two more principles arise from the transnational reach of these risks: inclusivity and sustainability. Governance must remain accessible to the entities least able to absorb compliance costs, and it must weigh the environmental and social impacts of large-scale deployment. Together, these requirements contribute to controllability, the ability to keep industrial AI under meaningful human and institutional command.

Drawn from engineering and safety science, controllability denotes the ability of human operators, regulators, and affected communities to understand, monitor, direct, and where necessary override the decisions of AI systems deployed in industrial settings. It is not a quality that a system either possesses or lacks, but a multidimensional continuum that must be assessed and governed across several institutional dimensions. The Controllability-Centred Governance Framework (CCGF) translates this single capacity into six core principles, presented here in the order in which they build upon one another: safety, reliability, transparency, accountability, inclusivity, and sustainability. Figure 2 presents the principles, and Table 2 summarises the governance requirements and the general standards that anchor each.

Safety is the first principle because it addresses the most direct consequence of cascading failure, namely physical and economic harm. It requires explicit safety objectives, systematic analysis of hazards, AI impact assessments throughout the AI lifecycle, and the reduction of residual risk to a level As Low As Reasonably Practicable (ALARP), documented through a structured safety, long used in aviation and the process industries. Under ISO/IEC 42001, AI impact assessment provides a systematic process for identifying and evaluating potential impacts, while ISO/IEC 42005 offers practical guidance for conducting and documenting such assessments.

Reliability extends this assurance across time. Where safety guarantees that a system is acceptable at the point of approval, reliability indicates whether it continues to perform under distributional shift, sensor noise, adversarial inputs, and gradual degradation – the conditions under which stealth risk arises. It therefore calls for performance benchmarks, continuous monitoring, and mandatory reporting of degradation. For this purpose, this paper proposes APR as a possible analytical tool – to be further tested and validated. The proposed graduated APRL scale is set out in Box 1 and described in detail in Annex A, including metrics and assurance procedures.

Transparency makes safety and reliability verifiable. It covers not only model explainability, but also technical documentation, training-data provenance, operational logging, and auditable records of human oversight, so that stealth risk and compound risk can be detected and their source traced back.

Accountability makes transparency actionable. It assigns clear responsibilities across developers, integrators, deployers, and operators. Effective accountability also depends on credible regulatory oversight and enforcement mechanisms capable of verifying compliance and ensuring that governance obligations are applied consistently across jurisdictions.

Inclusivity ensures that these obligations can be met by every participant in this shared system and not only by a limited number of firms in few economies. It calls for tiered compliance pathways, capacity building for the Global South, and safeguards against governance requirements becoming de facto trade barriers.

Sustainability completes the set by treating energy consumption, electronic waste, workforce displacement, and deskilling as design criteria rather than afterthoughts, so that the pursuit of near-term control does not accumulate longer-term costs that would erode the legitimacy and feasibility of the framework itself.

These six principles mutually reinforce one another rather than simply adding up, and they can be imagined as a chain in which each link depends on the one before it. Safety without reliability is incomplete, because a system judged acceptable at approval may not remain so in operation. Reliability without transparency cannot be verified, because sustained performance cannot be confirmed without documentation and logging. Transparency without accountability has little value,

since disclosure alone is insufficient if no party is responsible for what it reveals. Accountability without inclusivity excludes the smaller and developing-country producers most exposed to industrial AI risk yet least equipped to govern it. Finally, when pursued without taking sustainability into account, all five principles accumulate environmental and social costs that ultimately undermine the very controllability they were meant to secure. Controllability is, in this sense, a property that only emerges when all six principles are implemented together. Implementing these principles in practice requires governance arrangements that extend beyond individual organisations.

Effective regulatory oversight, international coordination, and capacity-building are essential to maintaining controllability across global industrial value chains. While no existing institution offers a complete model for AI governance, international experience with regimes that govern other safety-critical technologies shows that coordinated oversight, peer review, and technical cooperation can strengthen governance across borders. Such comparison must be drawn with care, since AI is more widely distributed, evolves more rapidly, and is more often privately developed and operated than the technologies those regimes were built to govern. Nonetheless, this experience offers useful lessons for developing internationally coherent approaches to industrial AI governance.

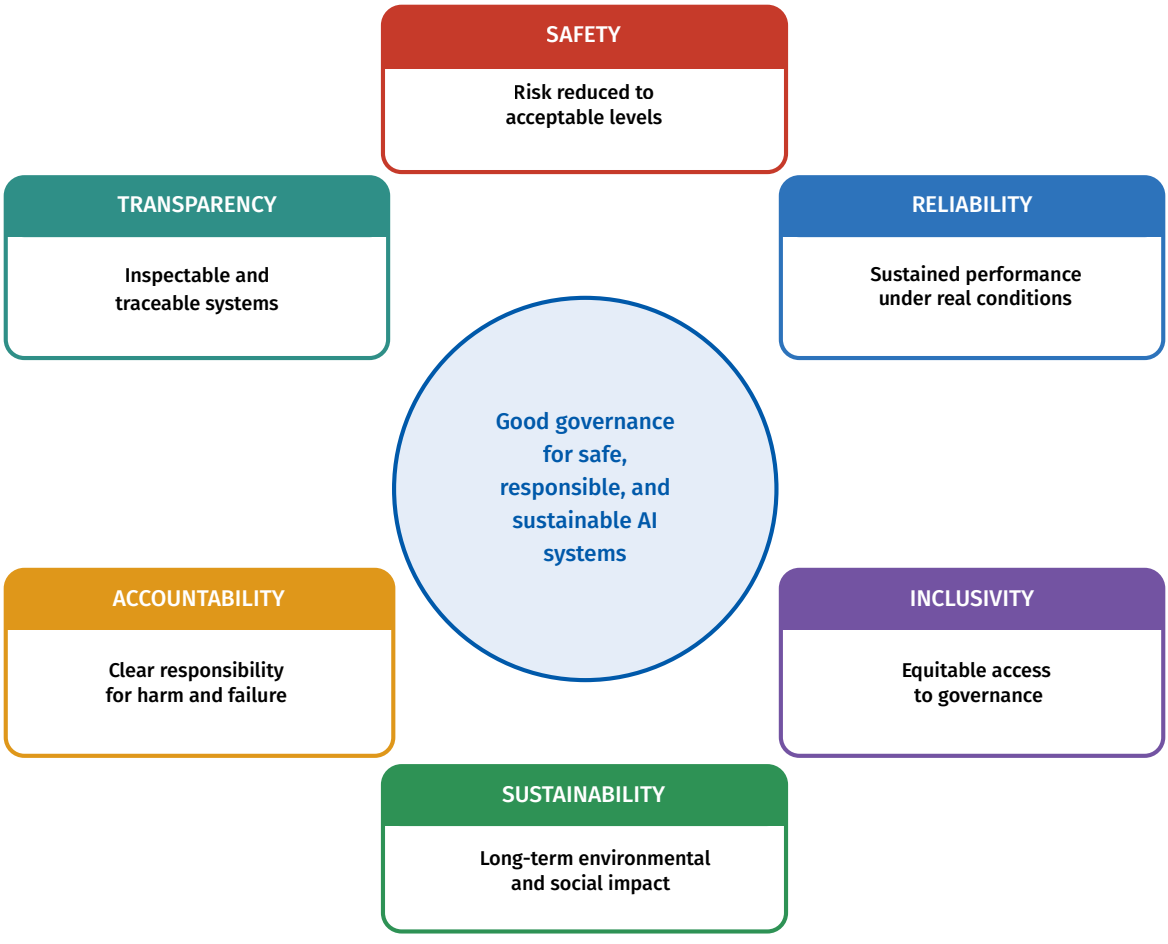


Figure 2. Six Core Principles of CCGF

On the Origin of AI Performance Readiness Levels (APRL)

APRL is an original construct introduced in this paper, modelled by analogy on the Technology Readiness Levels (TRL) scale developed for NASA by Mankins (1995) and since adopted across defence, space, and research-funding systems. Where Technology Readiness Levels (TRLs) assess technological maturity - from the observation of basic principles at TRL 1 to a system proven in an operational environment at TRL 9 - AI Performance Readiness Levels (APRLs) apply a comparable logic to operational AI reliability. APRLs progress from narrow validation in a laboratory setting at APRL 1 to sustained performance in full operational deployment under conditions of distributional shift at APRL 9.
APRL should not be confused with AI maturity or readiness tiers: it assesses the demonstrated reliability of an individual AI system in operation, whereas maturity and readiness scales - such as the shared five-tier reference within the wider United Nations system and the country-level readiness diagnostics in UNIDO's own suite - assess the AI capacity of countries or organisations. There is currently no harmonised readiness scale for operational AI reliability. APRL is described in detail in Annex A, which sets out its nine levels, seven measurement dimensions, and associated assurance procedures. Its empirical validation and calibration are identified as priorities for future research in Section 5.

Principle	Governance Requirement	Regulatory / Standards Anchor
Safety	Explicit safety objectives; systematic hazard analysis; ALARP residual-risk standard; safety-case documentation. AI impact assessments throughout the AI lifecycle.	EU AI Act; ISO/IEC 42001; ISO/IEC 42005; aviation and process-industry safety-case practice.
Reliability	Performance benchmarks; continuous monitoring; mandatory reporting of degradation; a proposed graduated APRL scale (see Box 1; full specification in Annex A).	ISO/IEC 42001; ISO/IEC 25059; Technology Readiness Levels analogue (Mankins, 1995; NASA, 2017).
Transparency	Technical documentation; training-data provenance; operational logging; human-oversight records.	EU AI Act; ISO/IEC 42001.
Accountability	Clear attribution across developers, integrators, deployers, and operators; non-enforceability of liability-diffusing disclaimers; regulatory oversight mechanisms and independent conformity assessment where appropriate.	ISO/IEC 42001; OECD AI Principles; EU AI Act.
Inclusivity	Tiered compliance pathways; technical assistance for SMEs and developing countries; anti-trade-barrier safeguards.	SDG 9; UNIDO ISID mandate; UNIDO policy research.
Sustainability	Integration of energy use, e-waste, workforce displacement, and deskilling as design criteria.	OECD AI Principles; UNIDO policy research (Lema, 2026).

Table 2. Six-Principle Matrix for the CCGF

2.3 From Algorithm Governance to System Governance

The six principles set out above do not apply to the algorithm alone. Controllability, safety, and reliability do not reside in a model considered in isolation; they result from how that model interacts with the sensors that feed it, the actuators it drives, the operators who supervise it, and the organisational routines and physical infrastructure within which it runs. Technical model governance, which covers training data, architecture, and output quality, is therefore necessary but not sufficient. For the six principles to be effective in practice, governance cannot stop at the algorithm; it must extend to the whole sociotechnical system in which industrial AI is embedded.

This marks a departure from the first wave of AI governance, which focused on the algorithm: training data, model architecture, output quality, and the direct impact of model outputs on individuals. That focus is still appropriate for many consumer-facing applications, where the model's output is itself where the harm occurs. It is inadequate for industrial AI, where harm arises less from a single output and more from how the model behaves once it is wired into a production system. Table 3 compares the two governance approaches across the dimensions that matter most for industrial deployment.

Industrial AI operates within complex, adaptive sociotechnical systems, and the risks that matter are properties of those systems rather than of a single model. Governance that considers only the algorithm will therefore overlook emergent behaviour at the system-level and that that no single component would exhibit on its own. A predictive-maintenance model, a process-control model, and a workforce-scheduling system may each perform well in isolation, yet their interaction can produce risk modes that none of them would display alone. System-level governance must address precisely these interactions, across four dimensions.

Dimension	Algorithm-Level Governance	System-Level Governance
Unit of analysis	Individual model and its training data.	Sociotechnical system: algorithms, sensors, actuators, operators, routines, and infrastructure.
Risk focus	Model output quality; individual impact.	Emergent system behaviour; cascading and compound risks.
Integration boundaries	Rarely addressed explicitly.	Governed via APIs, data pipelines, safety cases, and human-override authority.
Human oversight	Binary: in-the-loop or autonomous.	Tiered: pre-approval, post-hoc review, or continuous monitoring.
Adaptation and learning	Assumed static after deployment.	Ongoing performance monitoring and mandatory reassessment on drift.
Organizational scope	Engineering or data-science team.	Enterprise-wide: AI owners, cross-functional committees, audits (ISO/IEC 42001).

Table 3. From Algorithm-Level to System-Level Governance of Industrial AI

The first dimension is the set of integration boundaries where a model connects with the rest of the system: its APIs, data pipelines, and control interfaces. Because risk concentrates wherever one component passes control to another, these boundaries must be governed by explicit conditions on data exchange, on human-override authority, and on whether a model may carry out safety-critical functions at all. Safety cases for instance, structured documents long used in aviation and the process industries, provide a proven means of documenting and meeting those conditions.

The second dimension, closely related, is human oversight, which cannot be reduced to a binary choice between keeping a human in the loop or ceding full autonomy. Oversight should instead be adjusted according to consequences: some decisions must receive pre-approval, others warrant only post-hoc review, and others may proceed autonomously provided that they remain under continuous monitoring.

The third dimension concerns governance over time, through continuous learning during operation. Industrial AI seldom remains fixed once deployed, since adaptive systems retrain and may drift away from the conditions under which they were initially validated. Governance must therefore require reassessment whenever performance shifts are detected and keep an auditable record of each retraining cycle, so that a system's behaviour at any given time can be traced to a documented decision.

The fourth dimension is organisational, and it carries as much importance as any technical control, because responsibility that is unclear in practice cannot be enforced on paper. Deploying enterprises should appoint accountable owners for each AI system, establish cross-functional governance committees, maintain incident-response protocols, and audit their AI governance at regular intervals.

ISO/IEC 42001:2023 provides a comprehensive management-system template for these functions, which can integrate easily with existing quality, safety, and environmental management systems.

Taken together, these four dimensions show that moving from the algorithm to the system is not simply a broadening of scope. It represents a fundamental shift in how industrial AI risk is understood: because such risk is a property of sociotechnical systems rather than of individual models, governance instruments must be designed accordingly.

3 Reshaping the Governance Approach and Architecture

3.1 From Competition to Shared Responsibility

The underlying philosophy of most national AI strategies is to enhance national competitiveness. Their explicit or implicit aim is to deploy industrial AI faster and more effectively than competitors, and thereby to capture the economic and strategic gains that early advantage brings. This reasoning is in line with national interests, and the competitive impulse often accelerated innovation. As a foundation for governance, however, this approach might prove incomplete. A framework built on competition treats safety, reliability, and controllability as costs to be minimised rather than as public goods to be secured, and it measures success by the speed of deployment rather than by the integrity of the system into which AI is deployed. The result is, as increasingly visible today, a relatively fragmented landscape of regulatory gaps and inconsistent standards, in which the response to failure tends to be reactive and unclear in terms of liability. This calls for a reconsideration of the underlying logic of AI governance (Figure 3).

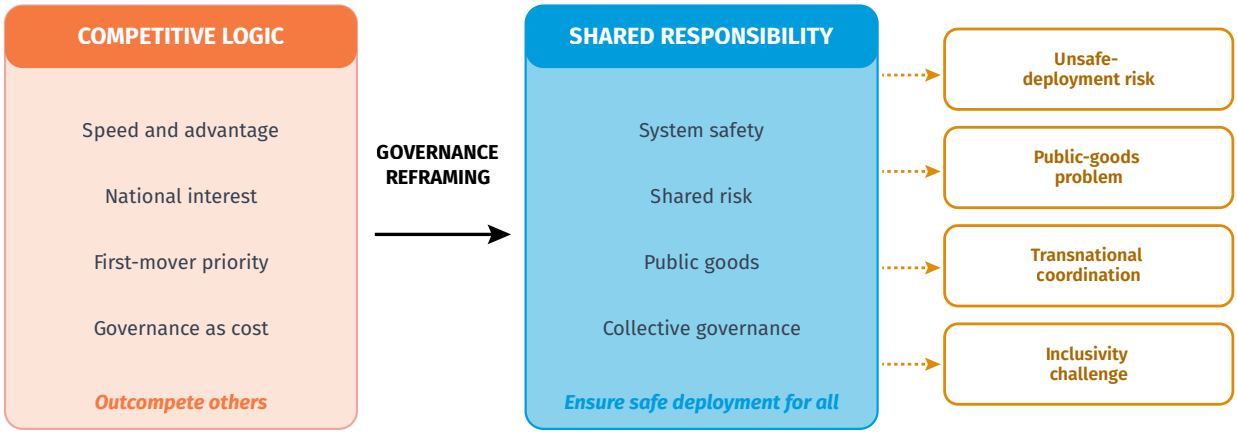


Figure 3. From Competition to Shared Responsibility

One possible alternative logic is that of shared responsibility. The idea draws on the disaster-risk governance literature (Tierney, 2014), which shows how those exposed to a common hazard are bound together by a collective obligation to manage it through cooperative institutions. Industrial AI gives rise to precisely such a community of shared exposure, because the failure of a single system can reach workers, firms, and communities far from the site where it is deployed. Applied to this setting, the idea reframes the central governance question: how to ensure that deployment across the global manufacturing system is safe, reliable, and controllable for every party that lives with its consequences, including the many who do not control the systems themselves? Framed in these terms, the shift is not a matter of preference but of necessity, and each of the implications that follow yields a governance advantage that competition alone cannot deliver.

The reframing begins by changing the understanding of where the main risk lies. Competition focuses attention on underemployment: the risk of falling behind, missing the productivity frontier, and ceding markets to faster movers. Shared responsibility identifies danger with unsafe deployment. Competitive pressure without a regulatory counterbalance rewards those who deploy before the infrastructure to contain failure is in place. Identifying unsafe deployment as the dominant risk is therefore a key advantage, since a framework can manage only the hazard it is built to recognise.

Because that hazard is shared, the qualities that contain it are shared as well, and this changes how they should be classified. Once exposure is held in common, safety, reliability, and controllability become public goods rather than private compliance costs. A firm that invests in them generates benefits that reach well beyond its own boundaries, including lower systemic risk, greater supply-chain resilience, and higher public trust. Yet it cannot fully capture the return on that investment. Because these benefits are non-excludable, purely market-based incentives will not provide enough. By classifying controllability as a public good, shared responsibility explains why governance cannot be left to the market and justifies instruments that align private incentives with the public interest: regulatory mandates, liability rules, procurement standards, and certification requirements.

Because the same hazard crosses borders, it also determines the level at which governance must operate. Industrial AI risk is transnational by nature: value chains are global, systems are developed in one jurisdiction and deployed in another, and a governance failure in one setting produces effects in many others. National frameworks designed in isolation will therefore leave structural gaps precisely where the risk concentrates. Competition, which treats other nations as rivals, does not seem to offer a sufficient remedy. Shared responsibility, which treats them as fellow parties exposed to a common hazard, reframes the core problem as one of transnational coordination and makes cooperative arrangements a requirement rather than an aspiration.

And because shared responsibility extends to all parties exposed to the hazard, it places developing countries at the heart of the design rather than at its margin. Under a competitive approach, they face a double challenge: the risk of being excluded from the norm-setting processes and being subjected to standards that do not reflect their industrial contexts. Shared responsibility, by contrast, requires an inclusive governance, with participation of developing countries in standard-setting, tiered compliance pathways that match obligations to capacity, and targeted investment in the public-good infrastructure on which capacity-building depends.

These implications reinforce one another, and together explain why this reframing is more than rhetorical. Shared responsibility does not displace national ambition; it provides the shared foundation without which that ambition would be self-defeating, since no manufacturer is secure within a system whose other nodes are unsafe. Its governance advantage is cumulative: it identifies the dominant risk, classifies the protective qualities as public goods, locates the problem where it actually concentrates, and accounts for every party the hazard reaches. Building on this, Section 3.2 tries to operationalise this new logic through a proposed three-level nested architecture, while Section 3.3 maps the governance instruments that would put the governance architecture into effect.

3.2 A Three-Level Nested Governance Architecture

Shared responsibility and the six core principles call for a governance architecture that no single institutional level can provide alone. The principles outline what responsible industrial AI requires, while shared responsibility establishes that these requirements are public goods whose impact is cross-organisational and cross-national. Securing them therefore requires action at several levels at once: the level where shared standards and cooperation are established, the level where binding rules are made and enforced, and the level where AI systems are actually deployed and operated. This paper accordingly proposes a three-level nested governance architecture spanning the global, national, and industry and enterprise levels. Figure 4 presents the architecture, and Table 4 sets out the core function of each level alongside its main instruments and actors.

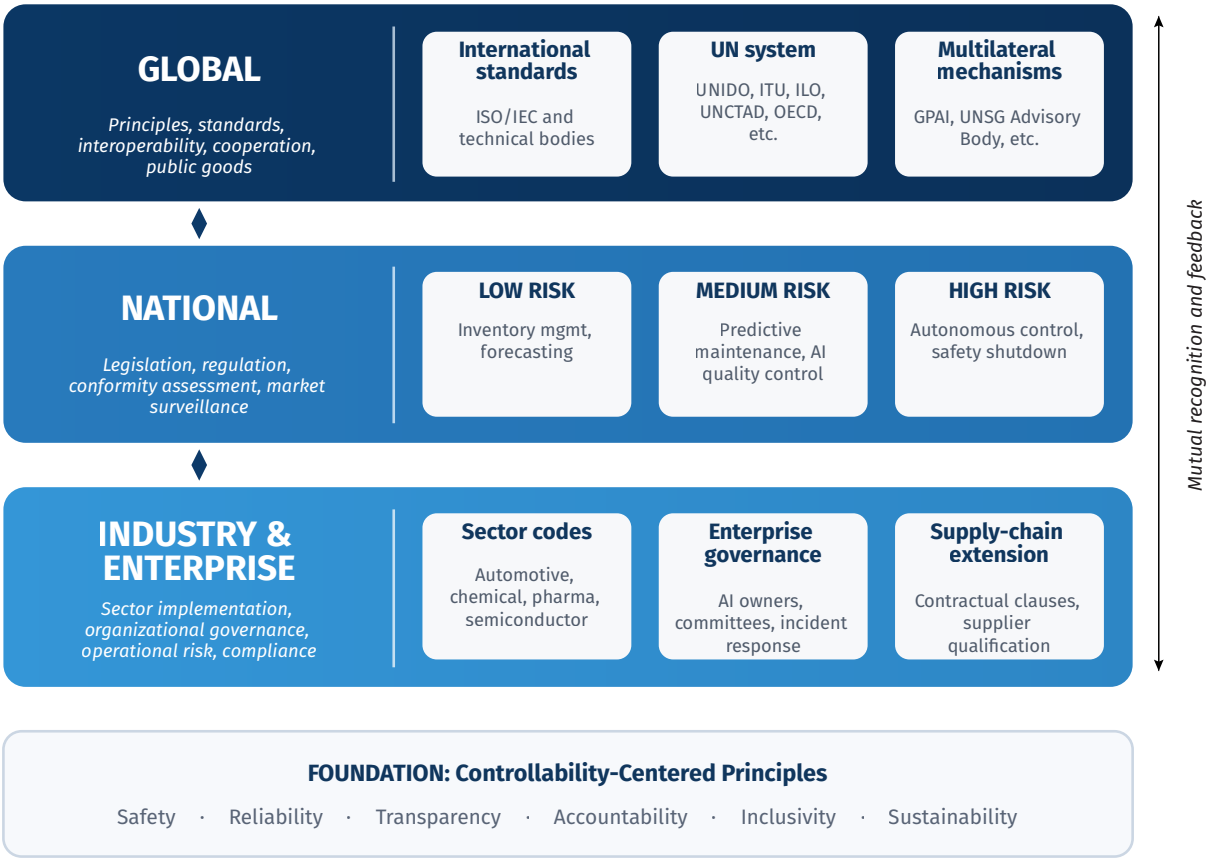


Figure 4. The Three-Level Nested Governance Architechture

The levels are nested rather than parallel. Each operates within the constraints set above it and supplies the operational substance that the level above cannot generate on its own: global principles acquire binding force only upon national ratification, and national regulations rely upon enterprise compliance. What holds the three together is a set of connective mechanisms. Mutual recognition allows a conformity decision taken in one jurisdiction to be accepted by others, which reduces duplication and lowers the compliance costs of firms that operate across borders. Information sharing, especially on incident intelligence and performance data, allows each level to learn from events observed by the others. Coordination keeps objectives aligned, so that global principles, national rules, and enterprise-level practices reinforce rather than contradict one another. Each level must therefore be coherent on its own terms, and the three together must create a mutually reinforcing mechanism that carries the six principles from shared norm to operational routine.

Level	Core Function	Key Instruments and Actors
Global	Principles, international standards, interoperability, cooperation, and global public goods.	Harmonised standards, mutual recognition of conformity assessment, and shared public goods such as incident-intelligence exchange, advanced through ISO/IEC, ITU, UNIDO, OECD, GPAI, and the United Nations Secretary-General's High-level Advisory Body on AI, etc. UNIDO's comparative advantage lies at the unique industrial application perspective of AI governance: manufacturing, quality infrastructure, standards implementation, conformity assessment, SME capacity-building, and developing-country industrial readiness.
National	Legislation, regulation, conformity assessment, and market surveillance.	Tiered risk classification, national AI registries, conformity-assessment regimes, and regulatory sandboxes, enacted by parliaments and applied by sectoral regulators and national standards bodies.
Industry and Enterprise	Sector-specific implementation, organisational governance, operational risk management, and compliance.	Sector codes, designated AI-system owners, internal audit, supplier qualification, and supply-chain clauses, carried by industry associations, deploying firms, auditors, and workers' representatives.

Table 4. Core Functions and Key Instruments Across the Three Governance Levels

3.2.1 Global Level: Standard Harmonisation and Cooperative Mechanisms

The global level is the appropriate level for tasks that cannot be discharged effectively at any lower level. It sets the baseline standards needed for cross-border coherence, supplies the public goods that no single nation has sufficient incentive to provide, coordinates jurisdictions on matters of shared concern, and creates equitable entry points for developing-country participation.

International standards are the primary instrument. The work of ISO/IEC JTC 1/SC 42 is a reference point: its AI standards, including ISO/IEC 42001, introduce a management-system approach to AI governance that gives organisations a structured basis for defining objectives, implementing controls, and pursuing continuous improvement. Because such standards are designed to be adaptable, they can serve as reference tools for adaptation to national and sector-specific contexts rather than as rigid prescriptions.

UNIDO's contribution at this level builds on its long-standing work on quality infrastructure and conformity-assessment systems in developing economies. It can engage with the international AI-standards process, including the AI management-system work of ISO/IEC, and can help assess how existing and emerging standards can be applied in industrial settings. Practical avenues for such cooperation include piloting standards in real industrial use cases through UNIDO-supported Centres of Excellence, aligning training and capacity-building with those standards, and developing small-scale testing environments for AI systems. Engagement of this kind also calls for coordination across related domains, since topics such as data governance and cross-border data flows fall under other parts of the international standards system.

UNIDO's distinctive role is not to duplicate standard-setting but to act as a bridge: ensuring that industrial AI governance reflects the needs and constraints of industrialising economies, and that its benefits reach firms in developing countries, particularly SMEs with limited governance capacity.

3.2.2 National Level: Tiered Regulation and Integrated Risk Management

If the global level establishes shared norms and standards, the national level is where these can be translated into binding requirements. National governance is the primary site of operationalisation, through legislation, regulation, conformity assessment, and market surveillance. Given differences in legal systems, administrative capacity, industrial structures, and governance maturity, the architecture defines international-level outcomes and principles while allowing national authorities discretion over implementation pathways.

Risk tiering is the central national instrument, because it makes governance proportionate to consequence. The framework distinguishes three tiers, set out in Table 5. National regulators can draw on criteria used in international practice, including the EU AI Act, and adapt sector-specific thresholds to national industrial conditions in consultation with industry and other affected stakeholders.

Tier	Application Context	Examples	Governance Obligations
Low	Non-safety-critical; failures readily detectable and reversible.	Inventory management; demand forecasting for non-critical supply-chain functions.	Transparency documentation; baseline performance monitoring.
Medium	Significant economic or operational harm on failure; human-in-the-loop.	Predictive maintenance with human oversight; AI-assisted quality control.	Conformity assessment; system registration; ongoing performance monitoring.
High	Safety-critical, autonomous, or integrated into critical infrastructure.	Autonomous process control in chemicals; AI safety-shutdown systems; AI-enabled industrial autonomous vehicles.	Mandatory conformity assessment; formal safety case; human oversight; registry entry.

Table 5. National Risk Tiering for Industrial AI

Tiering works best as an extension of arrangements that already exist rather than as a parallel regime. Many countries already regulate industrial process safety through mature instruments, such as the European Union's Seveso framework and the United States' OSHA process-safety management standard, with analogous arrangements in Japan and other Member States. Extending these regimes to cover AI risk is more efficient than building new institutions, and it draws on established enforcement mechanisms, inspectorate capacity, and industry familiarity.

Regulatory sandboxes complement this approach by generating evidence before full commercial deployment. By operating controlled environments in which firms deploy AI under supervision, national regulators can observe real-world performance and risk characteristics at manageable scale.

3.2.3 Industry and Enterprise Level: Operationalisation and Coordinated Constraints

At the industry and enterprise level, governance meets practice. Here the principles and structural requirements set above become organisational routines, technical implementations, and incident protocols.

Sector codes are the bridge between framework and practice. Developed by industry associations and sector regulators, they translate the CCGF into requirements specific to each industry's failure modes. The most natural early candidates are sectors that already possess a strong safety culture, established regulatory infrastructure, and well-defined operational standards, among them automotive, chemicals, pharmaceuticals, semiconductors, and electronics. Within these sectors, codes can specify risk-tiering criteria adapted to characteristic failure modes, human-oversight requirements for high-risk applications, performance benchmarks and monitoring protocols, and incident-reporting procedures. ISO/IEC 42001 supplies the management-system frame, and the sector codes supply the operational content that the frame leaves open.

Enterprise governance structures carry these requirements into the firm, and they should be proportionate to the risk profile of the systems deployed. Core elements include designated AI-system owners with documented responsibility, cross-functional governance committees, incident-response protocols for AI-related events, and regular internal audits of governance performance. Because the Plan-Do-Check-Act cycle of ISO/IEC 42001 mirrors the logic of existing quality, safety, and environmental management systems, enterprises can integrate AI governance into structures they already use rather than building separate apparatus.

3.3 Governance Instruments

The three-level architecture establishes where governance responsibility sits, but it acquires operational content only through the instruments that carry it. The architecture is operated through three instrument families, namely registration and disclosure, tiered risk classification, and enabling public goods, each anchored at a particular level of the architecture. Together they give the logic of shared responsibility its concrete machinery: they make industrial AI visible, render governance proportionate to consequence, and supply the common infrastructure that no single party will provide alone. The three also build on one another as a dependent sequence, in which each instrument rests on the information and capacity established by the one before it. Table 6 maps the instrument families to their governance functions, their primary level in the architecture, their design considerations, and their main implementing actors.

Registration and disclosure form the foundation of the toolkit, and they are primarily a national instrument. National AI registries and disclosure requirements provide regulators with information on high-risk industrial AI systems. Key information should cover system characteristics, risk tier, conformity-assessment status, designated ownership and relevant operational records. For the instrument to be accessible, its design must keep the compliance burden low, particularly for SMEs: this means deploying digital submission, risk-based scope and phased implementation. Shared content standards and interoperable formats can support information exchange and mutual recognition across jurisdictions. Appropriate disclosure should also provide workers, their representatives and supply-chain partners with information relevant to decisions and risks that affect them.

Instrument Family	Governance Function	Primary Level in the Architecture	Design Considerations	Main Implementing Actors
Registration and Disclosure	Provides the informational infrastructure on which other instruments depend.	National, enabled by global standards and interoperability.	Digital submission; tiered requirements; phased roll-out; disclosure to workers and supply-chain partners as well as regulators.	National AI authorities, supported by international organisations such as UNIDO.
Tiered Risk Classification	Operationalises governance proportionality; prevents regulatory arbitrage.	National, with criteria harmonised at the global level and applied at the industry and enterprise level.	Criteria harmonised across jurisdictions; reference to the EU AI Act.	National regulators; ISO/IEC coordination; industry associations and deploying firms.
Enabling Public Goods	Supplies non-excludable, non-rivalrous infrastructure that the market under-provides.	Global, with national governments and development finance institutions.	Testing platforms; benchmark datasets; incident databases; independent audit capacity; technical assistance.	National governments, supported by international organisations such as UNIDO and development finance institutions.

Table 6. Governance Instruments Mapped to Three-Level Architecture

Tiered risk classification builds directly on this informational foundation and converts registration information into proportionate obligations across all three levels. It is enacted as a national instrument: once the population of deployed systems has been registered and disclosed, national regulators can sort it into risk tiers and match the intensity of obligations to the severity of potential harm, concentrating scrutiny where consequences are gravest while sparing low-risk applications from disproportionate requirements. Yet proportionality is only half of the instrument's value, because tiering also encourages cross-border consistency. For mutual recognition to be feasible and regulatory arbitrage to be foreclosed, tiering criteria must be harmonised at the global level, so that risk classifications are well aligned. The CCGF tiering structure offers an initial basis for international agreement on common classification criteria. Tiering is ultimately applied at the industry and enterprise level, where sector codes translate the common criteria into thresholds calibrated to each industry's characteristic failure modes, and where deploying firms assign their systems to the appropriate tier and meet the applicable obligations.

Registration and tiering, in turn, requires supporting infrastructure, and such infrastructure is precisely what markets under-provide. Testing platforms, benchmark datasets, incident databases, independent audit capacity, and technical cooperation programmes meet the typical definition of public goods, since no single firm can capture their full value. Left to the market, they are systematically undersupplied, and the registration and tiering regimes that rely on them remain formal rather than effective. Their provision therefore requires coordinated contributions from national governments, and relevant international organisations. UNIDO could support and contribute to that effort, since its established role in industrial-development technical assistance, together with its institutional relationships with industrial-development bodies across developing countries, equips it to help build public-good infrastructure that reaches the firms most at risk of being left behind.

Taken together, registration and disclosure, tiered risk classification, and enabling public goods give the three-level architecture its working instruments, anchoring shared responsibility at the level where each can be carried out most effectively while binding the levels to one another through shared standards, harmonised criteria, and common infrastructure. How this governance design can be put into practice, sequenced over time and grounded in the widely differing conditions of UNIDO member states, is the task of the implementation pathways set out in Section 4.

4 Implementation Pathways

4.1 Phased Implementation Strategy

Moving from the current fragmented landscape to the three-level nested architecture cannot be accomplished in a single step. The transition calls for a phased strategy that remains realistic about institutional capacity, the current governance landscape, and the pace at which governance infrastructure can be built in practice. Given the diversity of contexts across Member States, the roadmap proposed below is intended as a general reference: it sequences a common set of building blocks rather than prescribing a single model, and each phase can be entered when the necessary conditions are in place. The three phases are overlapping rather than strictly sequential, allowing early movers to advance while others are still laying foundations, with each phase building on the one before. Figure 5 and Table 7 summarise the proposed sequence.

Implementation of the roadmap will require coordination among a range of stakeholders across the UN system and beyond. Existing initiatives and partnerships, including, for example the AI Governance for Humanity Lab[1] , may potentially provide valuable contributions in areas aligned with their respective mandates, including by supporting dialogue, knowledge exchange, capacity-building, and practical cooperation to strengthen the emerging global AI governance architecture. The specific roles of different actors will depend on evolving priorities, institutional capacities, and opportunities for collaboration.

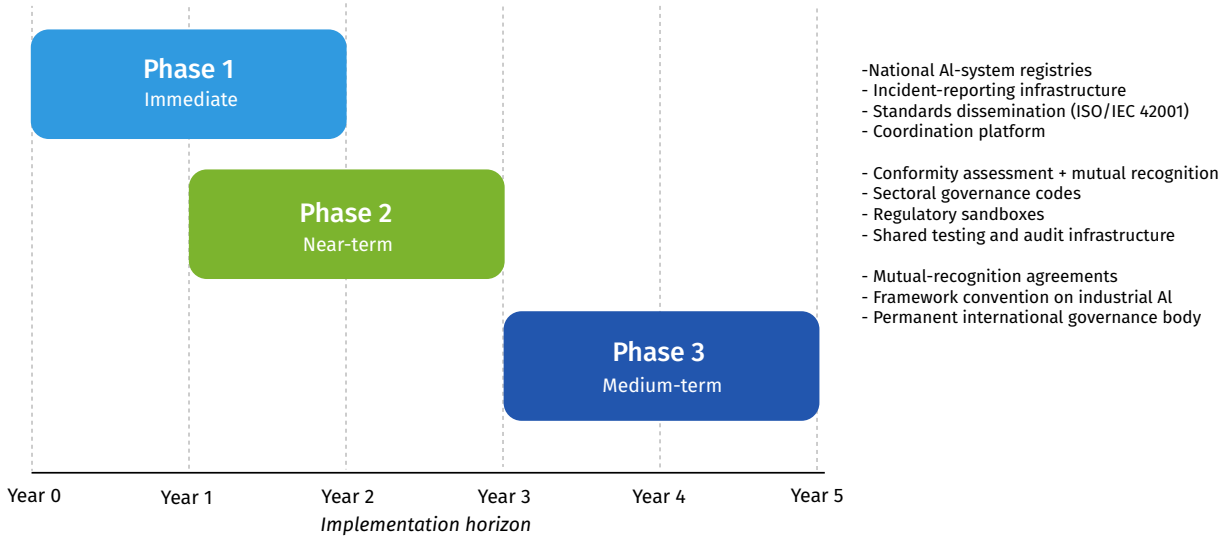


Figure 5. Phased Implementation Roadmap for Industrial AI Governanace

1. AI Governance for Humanity Lab. (2026). Private-sector AI Governance in practice: Insights, challenges and emerging cooperation approaches

Phase	Horizon	Priority Actions	Principal Actors
Phase 1. Immediate	Years 0-2	Establish national AI-system registries; build incident-reporting infrastructure; advance actions related to ISO/IEC 42001; and launch an international coordination platform. UNIDO could play a supporting and facilitating role, while the platform could draw on existing UN initiatives and multi-stakeholder partnerships. Relevant potential implementation partners - including, for example, the AI Governance for Humanity Lab - could contribute to knowledge exchange, multi-stakeholder dialogue, and practical cooperation, in line with their respective mandates and expertise.	National regulators; UNIDO; potentially AI Governance for Humanity Lab; standards bodies.
Phase 2. Near-Term	Years 1-3	Stand up conformity-assessment regimes with mutual recognition; publish sectoral governance codes; operationalise regulatory sandboxes; invest in shared testing and audit infrastructure.	National governments; industry associations; development finance institutions.
Phase 3. Medium-Term	Years 3-5	Negotiate mutual-recognition agreements; explore options for strengthened international cooperation, including, in the longer term, possible framework arrangements for industrial AI governance.	States as negotiating parties; United Nations system; UNIDO, together with relevant implementation partners including potentially the AI Governance for Humanity Lab, in a supporting role, subject to Member State guidance.

Table 7. Phased Implementation Strategy: Horizon, Actions, and Actors

Phase 1 occupies the immediate horizon of years 0 to 2 and lays the institutional and informational foundations on which all later development depends. Its first task is to make industrial AI legible to those who must govern it. National AI-system registries are therefore established for high-risk systems, recording for each entry the system description, application domain, risk tier, conformity-assessment status, designated owner, and incident contact. A registry captures only a static snapshot, however, so it is paired from the outset with mandatory incident reporting, which steadily builds the empirical base that current governance lacks. As this informational backbone takes shape, standards awareness and adoption are promoted through targeted dissemination of ISO/IEC 42001, supported by UNIDO and accompanied by technical assistance to SMEs and developing-country enterprises, so that the cost of early compliance does not fall hardest on those least able to bear it.

These national foundations are then connected at the global level by an international coordination platform, which UNIDO could support and facilitate, working with the AI Governance for Humanity Lab as a potential implementation partner to promote knowledge-sharing, convene multi-stakeholder dialogue and support practical cooperation, and which brings regulators, industry associations, standards bodies, and developing-country representatives into a common forum and helps ensure that the foundations laid in different jurisdictions are broadly compatible from the start rather than reconciled with difficulty later.

Phase 2 spans the near-term horizon of years 1 to 3 and builds on these foundations to put in place the operational machinery of systematic implementation. With registration and reporting established, attention turns to verification: national authorities set up or designate conformity-assessment regimes and connect them through mutual-recognition arrangements, so that a system assessed in one jurisdiction need not be assessed again in full elsewhere and cross-border compliance burden falls accordingly. This verification capacity is complemented at the industry and enterprise level by sector-specific governance codes, published in priority sectors such as automotive, chemicals, pharmaceuticals, and electronics, which translate the common principles into the operational language of each industry. To test such arrangements against reality before systems reach full commercial use, regulatory sandboxes come online for controlled deployment under supervision, generating evidence on real-world performance and risk. Because none of this can rest on capabilities that individual firms will finance on their own, national governments and international development partners invest during this phase in shared testing platforms, benchmark datasets, and independent audit capacity, the common infrastructure on which credible conformity assessment ultimately depends.

Phase 3 reaches the medium-term horizon of years 3 to 5 and completes the transnational layer that shared responsibility ultimately requires. By this stage the mutual-recognition arrangements developed in Phase 2 can mature into broader mutual-recognition agreements, allowing enterprises that have satisfied governance requirements in one jurisdiction to deploy in another without redundant conformity assessment.

To give these agreements a durable anchor, Member States may wish to explore options for strengthened international cooperation, including, in the longer term, possible framework arrangements for industrial AI governance. It would be for Member States to explore, through consultation and coordination, the institutional form of arrangements, including how functions such as coordinating global industrial AI governance, maintaining standards repositories, administering mutual-recognition agreements, and providing ongoing technical assistance would be supported. With this step, the global level would acquire a firmer institutional footing, so that the national and industry and enterprise levels can operate within a coherent and predictable international framework, and the nested architecture built up across the preceding phases is complete.

4.2 Grounding the Framework: Country-Level Opportunities and Challenges

The phased strategy and governance instruments presuppose enabling conditions that are unevenly distributed across countries; this section frames the resulting country-level opportunities and challenges, explains the contribution UNIDO could make in addressing them, and describes its programmatic response.

4.2.1 Country-Level Opportunities and Challenges

The phased roadmap in Section 4.1 and the instrument toolkit in Section 3.3 identify the actions needed to establish the proposed governance architecture. However, implementation will depend on countries having sufficient digital capabilities, data governance arrangements, and institutional coordination. National registries, conformity assessment systems, regulatory sandboxes, and shared testing infrastructure all require a minimum level of national capacity. Addressing these capacity constraints is therefore an integral part of the framework rather than a separate consideration. Each challenge outlined below also represents an opportunity where targeted support can accelerate implementation and strengthen long-term governance.

Many Member States have launched industrial AI pilot projects, but relatively few have been scaled up. The main challenge is not generating new ideas but creating the conditions for wider adoption. Pilot projects often remain isolated because they are not supported by common standards, interoperable data infrastructure, or appropriate financing mechanisms. As a result, lessons learned are not translated into broader industrial adoption or governance practice. The proposed architecture is intended to address these barriers by providing the institutional and technical foundations needed to support scaling and replication.

Foundational capability gaps further constrain the diffusion of successful pilots. Even where pilot projects are underway, their sustainability is often limited by weak product traceability systems, immature data management practices, and limited alignment with evolving international requirements such as ISO/IEC 42001 and sector-specific conformity standards. Evidence from the AI Governance for Humanity Lab Insight Report also shows that many companies across the Global South face significant digital capability and governance challenges, highlighting the broader capacity constraints that many developing countries face in creating an enabling environment for responsible industrial AI. These gaps become increasingly important as global value chain leaders introduce AI-related compliance requirements for their suppliers. Countries that cannot demonstrate a basic level of governance capability risk being excluded from higher-value segments of industrial production, while those that strengthen these capabilities early can improve their access to international markets.

This capability deficit has a structural counterpart in the gap between data, systems, and use cases. Industrial data is often siloed, unstructured, or inaccessible because operational-technology and information-technology environments remain fragmented, while legacy production systems resist integration and deployable, value-generating use cases are hard to identify. The result is an industrial AI bottleneck in which the technology's potential goes unrealised not for want of algorithms but because the system conditions for deployment have not been assembled. The difficulty is most acute for SMEs in developing countries, which seldom hold the in-house technical capacity to bridge it, and which therefore stand to gain most from shared infrastructure and external support.

These deficits converge at the level of the governance ecosystem. The instruments proposed in Section 3.3, namely registration and disclosure, tiered risk classification, and enabling public goods, each rest on a minimum of national data governance, interoperability, and trusted data-sharing arrangements. A national registry cannot be meaningfully populated where enterprises cannot generate conformity documentation, and risk tiering cannot be applied consistently without sector-specific benchmarks calibrated to national industrial contexts. Absent deliberate investment in ecosystem-level governance, the framework's instruments will not take root.

Considered as a whole, these constraints are real but addressable, and the entry point is diagnostic. Systematic country-level diagnosis, which maps digital maturity across industrial sectors, identifies the highest-leverage capability gaps, and sequences interventions, accordingly, converts a diffuse set of obstacles into an ordered work programme. UNIDO's country presence, industrial-development mandate, and partnerships with national counterparts equip it to support this function and to connect its findings to the implementation roadmap and instruments described in Sections 4.1 and 3.3. That diagnostic-and-delivery task is what the remainder of this section addresses.

4.2.2 Why UNIDO: A Comparative-Advantage Argument

The diagnostic-and-delivery task identified above calls for an institution with an unusual combination of attributes: a mandate centred on industrial development, a technical track record in the disciplines that AI governance now requires, working relationships with both international standards bodies and national authorities, and a standing neutral enough to be trusted across industrial blocs. UNIDO combines experience across all four.

The foundation is mandate. UNIDO is the specialised agency of the United Nations with the unique mandate to promote inclusive and sustainable industrial development (ISID)[2]. That mandate is reflected directly in SDG 9 on resilient infrastructure, inclusive and sustainable industrialisation, and innovation, for which UNIDO serves as custodian agency for several global SDG indicators[3]. The Organization can support the participation of industrialising economies in global AI governance discussions, a voice that remains systematically under-represented in AI fora. Inclusivity is therefore not an add-on to UNIDO's involvement; it is intrinsic to its mandate.

Mandate is matched by a relevant technical record. For decades UNIDO has provided technical expertise to enhance the quality infrastructure across the system of standards, metrology, accreditation and conformity assessment in Member States. It has supported more than 100 countries in developing quality-infrastructure systems and helped more than 600 laboratories achieve international accreditation[4], alongside a substantial active portfolio of quality and trade-related programmes, and it has developed instruments that help countries benchmark quality-infrastructure maturity and prioritise investment. This is precisely the institutional capability that industrial AI governance demands, because conformity assessment, accreditation, registries, and testing are the operational core of the toolkit set out in this paper: not novel functions but established ones applied to a new class of technology.

2. UNIDO. Who We Are.
3. United Nations Statistics Division. SDG Indicators Database; see also UNIDO, Sustainable Development Goals.
4. UNIDO Knowledge Hub - LABNET: A worldwide Laboratory Network.

A third advantage lies in partnerships. UNIDO has cooperated with the International Organization for Standardisation for more than four decades, a collaboration formalised through a memorandum of understanding and reaffirmed in recent years around the adoption of international standards aligned with SDG 9 and the empowerment of national standards bodies. It participates in the established network of international quality-infrastructure organisations alongside the main metrology, accreditation, standardisation, and conformity-assessment bodies, and it works closely with national standards bodies in international standard-setting and builds compliance capacity. These partnerships, from the global standards system to industrial-development counterparts in member states, allow UNIDO to translate global norms into national and industry practice. Its role is thus to bridge global standards and their application in industrial contexts.

The fourth advantage is standing. UNIDO acts as a neutral, impartial platform that delivers technical cooperation on a demand-driven basis among governments, the private sector, and other partners. UNIDO is mandated to serve industrialising economies on equal terms. Its function is advisory and capacity-building, helping Member States adopt international standards and build the institutions needed to implement them. This unique role enables UNIDO to contribute credibly to the provision of the public goods that a shared-responsibility approach requires, including testing platforms, benchmark datasets, incident databases, and technical assistance.

These advantages summarise the contributions UNIDO could make across the phased roadmap. In the near term it could serve in a supporting and facilitating role for the international coordination platform and, throughout the early phases, as a provider of technical assistance to SMEs and developing- industries. Over the longer term, should Member States decide to pursue framework arrangements for industrial AI governance, UNIDO could, subject to Member State endorsement, provide institutional support, governance arrangements appropriate to such a mechanism, and a sustainable financing model.

4.2.3 What UNIDO Can Do: Programmatic Response

UNIDO's comparative advantage is already expressed in programmes that operate at country level. The main delivery platform is the Global Alliance on Artificial Intelligence for Industry and Manufacturing (AIM Global), a multi-stakeholder alliance launched by UNIDO in 2023 to bring government, industry, academia, and civil society together around advanced, AI-driven manufacturing and fairer access to its benefits, particularly for developing countries. Working through a community and matchmaking platform and a growing network of Centres of Excellence (CoE), the initiative mobilises public and private partners, supports national AI and digital-transformation strategies, and supports applied use cases in priority sectors such as predictive maintenance, quality inspection, process optimisation, and supply-chain analytics. The CoE network, anchored by a flagship centre and extending through regional centres and delivery models that channel tested solutions and trained staff to smaller manufacturers, generates the implementation experience and local capacity that Phase 1 of the roadmap requires.

Programmatic delivery is reinforced by normative and policy instruments. UNIDO's policy research on responsible industrial AI, including its work on bridging the AI divide for developing-country manufacturers, articulates the Organization's position on responsible deployment. This body of work aligns with the six principles of the CCGF and gives national counterparts a structured reference for policy dialogue, investment prioritisation, and engagement with international standard-setting. It also serves a signalling function, communicating to member states and development partners that investment in governance is an industrial-development imperative rather than a compliance overhead.

A further line of work targets the ecosystem deficits identified above. UNIDO can help member states build the data-governance and infrastructure preconditions for effective AI deployment, including interoperable data frameworks, trusted data-sharing arrangements within and across sectors, and alignment with international standards such as ISO/IEC 42001. This support addresses the ecosystem-level deficits set out in Section 4.2.1 and builds the enabling environment that the Phase 1 registration and disclosure regimes need in order to function.

Binding these elements together is country diagnostics. UNIDO's country-level diagnostic work maps digital maturity across industrial sectors, identifies sector-specific entry points for AI deployment, and generates the evidence base for sequenced, context-sensitive intervention. That output feeds directly into the phased strategy: readiness assessments inform the design of Phase 1 registries, calibrate the risk-tiering thresholds applied at the national level, and guide the prioritisation of technical assistance for SMEs and industrial clusters. In this way the diagnostic closes the loop between the three-level nested architecture and the specific conditions under which it must work. This function is operationalised in UNIDO's AI and Digital for Industry Navigator (AIDIN), which aggregates indicators drawn largely from public international sources across five enabling pillars, namely infrastructure, data, the innovation ecosystem, capital, and governance, together with measures of industrial digitalisation and AI adoption, to generate a national AI and digital maturity profile and a set of tailored industrial transformation pathways. Where this paper sets out a normative governance architecture organised by level, AIDIN works at the level of country readiness, helping each country locate its starting point and sequence the steps toward it.

Country diagnostics are most effective when matched by a diagnostic at the level of the firm, because national readiness is ultimately realised in the operational capabilities of individual enterprises, and it is here that small and medium-sized enterprises most often fall behind. For this level, UNIDO's programmes can draw on tools such as the Operations Excellence Readiness Index (OPERI), an enterprise-level self-assessment developed by the International Centre for Industrial Transformation (INCIT), which scores a firm's operational and digital maturity across three building blocks, six pillars, and eighteen dimensions, complemented by productivity indicators such as labour productivity and capacity utilisation. It returns a benchmarked maturity rating and a prioritised set of improvement steps, giving a smaller manufacturer a concrete and affordable path from its current operations toward the readiness that responsible AI adoption presupposes.

Used together, the two instruments form a two-level diagnostic stack that mirrors the governance architecture of this paper. AIDIN assesses whether the enabling conditions for AI are in place at the national level, while OPERI assesses whether an individual enterprise is operationally ready to adopt it; the national profile identifies where capacity-building is needed, and the enterprise assessment directs it to the firms and the dimensions where it will have most effect.

Together with the proposed AI Performance Readiness Levels of Annex A, they would also complete a coherent set of readiness measures spanning the levels of governance: AIDIN gauges the readiness of the country, OPERI the readiness of the firm, and APRL the demonstrated reliability of the individual AI system. For small and medium-sized enterprises, and for developing economies in particular, this layered diagnosis converts a general commitment to inclusive participation into specific, sequenced, and fundable action, which is the practical substance of capacity-building under the framework.

Instrument	Governance level	Unit assessed	What it measures	Output / next step
AIDIN	National	Country	Enabling conditions across five pillars: infrastructure, data, innovation ecosystem, capital, and governance, with industrial digitalisation and AI adoption	National maturity profile and tailored industrial transformation pathways
OPERI	Industry and enterprise	Firm, especially MSMEs	Operational and digital maturity across three building blocks, six pillars, and eighteen dimensions, with productivity indicators	Benchmarked maturity rating and prioritised improvement steps
APRL (proposed; Annex A)	Industry	Individual AI system	Demonstrated operational reliability on a nine-level scale within a declared operating domain	Verified readiness level and the assurance required to deploy

Table 8. Layered readiness diagnostics, from national conditions to system reliability: UNIDO's AIDIN and OPERI instruments and the proposed APRL scale.

The analysis in this section converges on a concise set of priorities for the main actors:

- For UNIDO member states: mandate registration of high-risk industrial AI systems; actions related to ISO/IEC 42001; and integrate AI risk into existing process-safety regimes rather than creating parallel structures.
- For industry: designate accountable AI-system owners; adopt the six principles of the CCGF in enterprise policy; and extend AI risk-management clauses through supplier contracts.
- For the international community: negotiate mutual recognition of conformity-assessment results; establish an incident-intelligence exchange; and explore options for strengthened international cooperation, including, in the longer term, possible framework arrangements for industrial AI governance.
- For UNIDO: support and facilitate the international coordination platform in Phase 1 through AIM Global; deliver targeted technical cooperation through the Centres of Excellence network and country- and firm-level diagnostics (AIDIN and OPERI); and contribute to the provision of public-good infrastructure for testing, benchmark datasets, and incident databases, in line with Member State guidance, available resources, and multilateral coordination.

5 Conclusion and Areas for Further Work

Industrial including manufacturing AI has crossed a critical threshold. It is no longer an emerging technology whose governance can be deferred, but an operational reality whose risk modes are already materialising across global manufacturing. The governance response to date has not been adequate to the scale, character, and transnational reach of the risks involved. This paper has proposed a fundamental reorientation on three fronts: from competition to shared responsibility, from algorithm-level governance to system-level governance, and from single-level regulation to a three-level nested architecture spanning the global, national, and industry and enterprise levels. The Controllability-Centred Governance Framework and its six principles supply the principled foundation for that reorientation; the three-level nested architecture and its governance instruments give that foundation operational form; and the implementation pathways translate the resulting design into a realistic, phased route that can be entered as institutional capacity allows.

The benefits of this approach are not confined to a specific group of economies. For advanced economies, the architecture narrows the gap between the speed of deployment and the capacity to govern it. For developing countries, it opens a pathway to inclusive AI industrialisation that avoids the governance vacuum which would otherwise expose them to the full range of industrial AI failure risks without institutional protection. UNIDO's role follows from its mandate and its comparative advantage: it could support and facilitate the international coordination platform, provide technical cooperation and, over time and subject to Member State endorsement, contribute to the provision of the shared public goods and to any strengthened international cooperation arrangements that Member States may decide to establish.

A key conceptual contribution of this paper warrants explicit reflection in closing: the AI Performance Readiness Levels (APRL). Introduced in Section 2 and modelled by analogy on the Technology Readiness Levels long used, the APRL scale offers a graduated and shared measure of operational AI reliability, grading the demonstrated reliability of an individual AI system from narrow validation in the laboratory to sustained performance under real-world distributional shift. Its value to a shared-responsibility approach is direct: a common, graduated measure gives developers, deployers, regulators, and the communities exposed to industrial AI a shared reference for how reliable a system is, in place of the non-comparable assurances that prevail without a common standard. This paper does not leave APRL as an aspiration. Annex A specifies the scale in full, defining its nine levels in three assurance bands, the reliability dimensions along which systems are assessed, the metrics and evidence each level requires, and the assurance procedures by which a level is verified and maintained. What remains is empirical: translating the specified criteria into calibrated,

sector-specific thresholds and validating the scale against operational experience. That programme of field validation and calibration is the most immediate priority on the research agenda the paper now sets out.

Beyond the field validation of APRL, further work is needed to develop the conformity-assessment methods, benchmark datasets, and incident taxonomies on which the framework's instruments depend, to test sector codes against real failure modes in priority industries, and to refine country-level diagnostics so that the architecture can be calibrated to widely differing national conditions. This agenda is complementary to the evidence-first orientation of the Independent International Scientific Panel on Artificial Intelligence (2026), which identifies reliability mechanisms as a governance priority but does not itself supply a shared measure of operational reliability; APRL is offered as exactly such a measure for the industrial domain. Advancing this agenda is itself a shared undertaking and pursuing it in common is the surest way to ensure that industrial AI develops as a safe, reliable, and controllable foundation for inclusive and sustainable industrial development.

References

i. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.
ii. Independent International Scientific Panel on Artificial Intelligence. (2026). Preliminary Report: A Global Scientific Assessment of the Capabilities, Opportunities, and Risks of Artificial Intelligence. New York: United Nations.
iii. International Organization for Standardization and International Electrotechnical Commission. (2023). ISO/IEC 23894:2023, Information Technology, Artificial Intelligence, Guidance on Risk Management. Geneva: ISO/IEC.
iv. International Organization for Standardization and International Electrotechnical Commission. (2023). ISO/IEC TR 24029-2:2023, Artificial Intelligence, Assessment of the Robustness of Neural Networks, Part 2: Methodology for the Use of Formal Methods. Geneva: ISO/IEC.
v. International Organization for Standardization and International Electrotechnical Commission. (2023). ISO/IEC 25059:2023, Software Engineering, Systems and Software Quality Requirements and Evaluation (SQuaRE), Quality Model for AI Systems. Geneva: ISO/IEC.
vi. International Organization for Standardization and International Electrotechnical Commission. (2023). ISO/IEC 42001:2023, Information Technology, Artificial Intelligence, Management System. Geneva: ISO/IEC.
vii. International Telecommunication Union. (2024). AI for Good Platform. Geneva: International Telecommunication Union.
viii. Labrunie, M., Fan, Z., and Leal-Ayala, D. (2026). IID Policy Brief No. 30: AI and the Future of Industry. Vienna: UNIDO.
ix. Lema, R. (2026). IID Policy Brief No. 29: The Green Transformation and the Future of Manufacturing. Vienna: UNIDO.
x. Mankins, J. C. (1995). Technology Readiness Levels: A White Paper. Washington, DC: NASA Office of Space Access and Technology.
xi. National Aeronautics and Space Administration. (2017). Technology Readiness Level Definitions. Washington, DC: NASA.
xii. Organisation for Economic Co-operation and Development. (2019, revised 2024). Recommendation of the Council on Artificial Intelligence (OECD AI Principles). Paris: OECD.
xiii. Tierney, K. J. (2014). The Social Roots of Risk: Producing Disasters, Promoting Resilience. Stanford: Stanford University Press.
xiv. United Nations. (2025). Innovative Voluntary Financing Options for Artificial Intelligence Capacity-Building: Report of the Secretary-General (A/79/966). New York: United Nations.
xv. United Nations Industrial Development Organization. (2024). Bridging the AI Divide: Empowering Developing Countries through Manufacturing. Vienna: UNIDO.
xvi. United Nations Industrial Development Organization. (2023). Industrial Development Report Series: Technology and Innovation for Sustainable Industrialization. Vienna: UNIDO.
xvii. United Nations Secretary-General's Advisory Body on Artificial Intelligence. (2024). Governing AI for Humanity. New York: United Nations.

Annex A

AI Performance Readiness Levels (APRL): Definitions, Metrics, and Assurance Procedures

This annex specifies the AI Performance Readiness Levels (APRL) introduced in Section 2 and Box 1. APRL is presented as a proposed analytical tool for further testing and validation, not as a standard or a formal UNIDO methodology; the specification that follows is the basis on which the scale can be tested, refined, and validated. It sets out what the scale measures, the dimensions along which reliability is assessed, the nine levels and the conditions each requires, the metrics and evidence on which a level rests, and the assurance procedures by which a level is verified and maintained. It then defines a protocol for validating the scale itself and illustrates the whole with a worked example. The specification is deliberately generic. It fixes the structure of the scale and the categories of evidence, while the numerical thresholds are set by sector, so that a single, comparable scale can be applied proportionately across industries and across levels of development without becoming a barrier to trade.

A.1 What APRL Measures, and What It Does Not

APRL grades the demonstrated operational reliability of an individual AI system within a declared operating domain, that is, the equipment, data conditions, and operating envelope for which the system has been validated. A reliability claim is always made relative to this domain. A system rated at a given level within one domain carries no rating outside it, and extending the domain requires fresh evidence.

APRL is distinct from two neighbouring ideas with which it is easily confused. It is not a measure of organisational or national maturity. Scales such as UNIDO's AIDIN diagnostics assess whether the enabling conditions for AI are in place across a country or an enterprise, whereas APRL assesses a single deployed system. Nor is APRL a risk classification. The national risk tiers described in Section 3 sort systems by the potential severity of their consequences, whereas APRL measures how reliably a given system actually performs. Risk classification asks how much reliability a system must demonstrate; APRL measures how much it has demonstrated. The two are joined in Section A.7.

A.2 Conceptual Basis, and One Deliberate Departure from TRL

As set out in Box 1, APRL is modelled on the Technology Readiness Levels developed for NASA (Mankins, 1995), transposing the logic of graduated readiness from technology maturity to operational reliability. It departs from that model in one respect that is essential to its purpose. Technology readiness is monotonic: once attained, maturity is not lost. Reliability is not like this. An AI system that performs well at the point of approval can degrade silently as operating conditions drift away from those under which it was validated, which is precisely the stealth risk the framework

is designed to catch. APRL is therefore a maintained grade rather than a one-time gate. A level reflects reliability that is currently demonstrated and continuously monitored. Following verified performance degradation, an APRL rating may be temporarily suspended or revised pending re-validation, and it is restored only when the evidence again supports it. This property, referred to here as revocability, is what allows the scale to track reliability across time rather than certify it once. It operates through defined monitoring thresholds and governance procedures rather than as an automatic or instantaneous downgrade.

A.3 The Dimensions of Reliability

APRL assesses reliability along seven dimensions. Five are the reliability sub-characteristics of the international AI quality model (ISO/IEC 25059): maturity, the breadth and rigour of validation relative to the operating domain; availability, the continuity of performant operation; fault tolerance, safe and graceful behaviour when components or inputs fail; recoverability, the capacity to detect, contain, and recover from failure; and robustness, the retention of performance under perturbation, adversarial input, and distributional shift, assessed following ISO/IEC TR 24029. To these the Controllability-Centred Governance Framework adds two dimensions without which industrial reliability cannot be governed: human oversight and controllability, the demonstrated capacity of operators to monitor, direct, and where necessary override the system; and monitoring and drift management, the presence and effectiveness of continuous performance monitoring, drift detection, and degradation reporting. A level is reached only when every dimension meets that level's requirement. Reliability is treated as a profile that must hold across all dimensions, not a single quantity that can be traded off among them.

A.4 The Nine Levels in Three Bands

The nine levels are grouped into three bands that correspond to the stages of a system's life and to the assurance appropriate to each. Table A1 defines the levels and the assurance mode attached to each; the paragraphs below give the intent of each band.

Band 1, Laboratory Validation (APRL 1 to 3), covers development, and a system in this band has not yet earned operational reliance. Its evidence is generated in controlled settings and rests on documented self-declaration. The band moves from a defined set of reliability requirements, a bounded operating domain, and a hazard analysis consistent with ISO/IEC 23894 (APRL 1), through validated performance on curated held-out data (APRL 2), to demonstrated performance retention across a defined battery of stress conditions, with failure modes catalogued and a safe fallback specified (APRL 3).

Band 2, Controlled and Pilot Deployment (APRL 4 to 6), covers supervised operation. The system runs in a real or representative environment under human oversight and live monitoring (APRL 4), sustains performance over time with working drift detection and at least one exercised recovery (APRL 5), and reaches a controlled operational release in which residual risk is reduced to ALARP and documented in a structured assurance case that is independently verified (APRL 6). Independent, third-party verification enters the scale at APRL 6.

Band 3, Operational Reliability (APRL 7 to 9), covers trusted operation at scale. The system is proven across its full operating domain over a defined observation period and entered in the national registry (APRL 7), maintains performance through actual distributional shifts and rare or adversarial events without safety-critical failure over an extended period (APRL 8), and finally sustains long-run reliability across contexts through a mature and continuously operating regime of monitoring, scheduled re-validation, and closed-loop incident learning (APRL 9). Systems at this band would be eligible for the mutual-recognition arrangements proposed in Section 3.

Band	APRL	Level	What must be demonstrated	Assurance mode
Band 1: Laboratory Validation (APRL 1 to 3)	1	Requirements and domain defined	Reliability targets, the operating domain, and main hazards are specified.	Self-declaration (documented)
	2	Validated in-distribution	Targets met on curated held-out data representative of the domain; baseline robustness probes run.	Self-declaration (documented)
	3	Stress-validated	Performance retained across a defined shift, edge-case, and adversarial battery; failure modes catalogued; safe fallback specified.	Self-declaration; optional accredited-lab review
Band 2: Controlled and Pilot Deployment (APRL 4 to 6)	4	Supervised pilot	Operates in a real or representative environment under human oversight and live monitoring.	Independent internal audit
	5	Sustained pilot	Performance held over time and varying conditions; drift detection active; recovery or rollback exercised.	Conformity self-assessment plus internal audit
	6	Controlled release	Residual risk reduced to ALARP; structured assurance case independently verified.	Third-party verification of the assurance case
Band 3: Operational Reliability (APRL 7 to 9)	7	Proven in operation	Sustained in-threshold performance across the full domain over a defined period; registered.	Accredited third-party audit
	8	Robust under real shift	Performance maintained through actual shifts and rare or adversarial events; no safety-critical failure over an extended period.	Accredited audit plus continuous surveillance
	9	Sustained and self-correcting	Long-run, cross-context reliability with continuous monitoring, scheduled re-validation, and closed-loop incident learning.	Continuous surveillance plus periodic recertification

Table A1. The nine AI Performance Readiness Levels, grouped into three assurance bands.

A.5 Metrics and Evidence

For each dimension APRL specifies a family of metrics rather than a single fixed number, because the value that counts as reliable differs by sector and by the severity of the application. Table A2 gives the indicative metric family and the primary evidence for each dimension. Two features of this approach follow from the framework's principles. First, thresholds are set by sector codes rather than centrally, so that a demanding safety-critical threshold in one industry does not become an inadvertent trade barrier in another: the scale is common, while the thresholds are proportionate. Second, evidence requirements escalate with the band, so that the cost of assurance is matched to the reliance placed on the system, and smaller producers operating at lower bands are not asked to bear the assurance burden of critical deployments.

Reliability dimension	Indicative metric family	Primary evidence
Maturity (validation breadth)	Coverage of the operating domain by validation scenarios; representativeness of test data	Validation protocol and coverage record
Availability	Operational uptime; share of operating time with performance above threshold	Monitoring logs
Fault tolerance	Share of induced faults handled by safe fallback; time to reach a safe state	Fault-injection results
Recoverability	Mean time to detect and to recover from degradation; rollback success rate	Incident and recovery records
Robustness	Performance retention under a defined perturbation, adversarial, and shift set (ISO/IEC TR 24029)	Stress-test matrix and results
Human oversight and controllability	Override coverage and latency; effectiveness of oversight where required	Oversight and override logs
Monitoring and drift management	Drift-detection latency; share of key indicators monitored; timeliness of degradation reporting	Monitoring configuration and drift logs

Table A2. Measurement dimensions, indicative metrics, and evidence

A.6 Assurance Procedures

Assurance escalates across the three bands and reuses the conformity-assessment machinery that the framework already establishes, rather than creating a parallel regime. In Band 1 a level is claimed by documented self-declaration, optionally reviewed by an accredited laboratory. In Band 2 the claim is supported by conformity self-assessment against the applicable sector code and by independent internal audit, and at APRL 6 by third-party verification of the assurance case. In Band 3 a level requires accredited third-party audit, continuous surveillance, and periodic recertification, together with entry in the national AI registry and, at APRL 9, prospective eligibility for mutual

recognition across jurisdictions. Because assurance rests on accreditation, conformity assessment, and surveillance, it draws directly on UNIDO's established quality-infrastructure expertise, and it offers producers in developing economies a graduated and affordable path into international markets in place of a single high barrier.

A.7 Linking APRL to the Risk Tiers

APRL and the national risk tiers of Section 3 are joined by a simple rule: before a system may be deployed, its risk tier sets both the minimum APRL it must hold and the assurance mode by which that level must be verified. Table A3 gives illustrative mapping. Higher-consequence applications require both a higher demonstrated reliability and more independent verification, while lower-consequence applications are held to proportionate, lighter requirements. The mapping is illustrative and is to be calibrated by each jurisdiction and sector. Its purpose is to show how a measure of demonstrated reliability turns a risk classification into an operational condition of deployment, closing the gap between how much reliability a system must show and how much it has shown.

Risk tier (Section 3)	Minimum APRL for deployment	Assurance mode	Re-assessment
Low	APRL 4	Documented self-declaration; baseline monitoring	On material change to the system or its operating domain
Medium	APRL 6	Conformity self-assessment; third-party verification of the assurance case; registration	Periodic, and on detected drift
High	APRL 7	Accredited third-party audit; formal safety case; continuous surveillance; registry entry (APRL 8 for autonomous safety-critical systems)	Continuous surveillance and periodic recertification

Table A3. Illustrative mapping of national risk tiers to a minimum required APRL and assurance mode.

A.8 Validating the APRL Construct

Two senses of validation must be distinguished. Using APRL to validate a system is the business of Sections A.4 to A.6. Validating APRL itself, that is, establishing that the scale measures what it claims and does so consistently, is a research task, and it is the most immediate priority on the agenda of Section 5. It proceeds in four stages, each with an acceptance criterion, and it is governed as a continuing programme rather than a single exercise.

Content validity. A structured expert elicitation, conducted as a Delphi exercise with AIM Global members, standards bodies, and sector regulators, confirms that the levels and dimensions are complete, correctly ordered, and free of gaps. The acceptance criterion is convergence of expert judgement across successive rounds.

Inter-rater reliability. Independent assessors grade the same set of systems against the rubric, and the rubric is refined until their agreement is acceptable. The acceptance criterion is a pre-specified level of inter-rater agreement, for example a weighted-kappa threshold agreed by the technical working group.

Predictive validity. Across a reference set of deployed systems, the APRL held at deployment is tested against subsequent operational outcomes. The acceptance criterion is that a higher APRL is associated with lower rates of incidents and unplanned degradation.

Calibration. Sector thresholds are tuned so that the same level denotes comparable reliability across industries. The acceptance criterion is confirmed cross-sector comparability on the reference set.

The construct is itself designed to be governed: it is proposed that APRL be maintained, versioned, and periodically revised by a technical working group, which UNIDO could help convene, subject to Member State guidance, so that the scale evolves with the technology and with accumulated operational evidence. This programme is what turns APRL from a defined scale into a validated one, and it is the sense in which the paper commits not only to defining the scale but to validating it.

A.9 A Worked Example: Predictive Maintenance

To show the scale in use, consider a predictive-maintenance system that forecasts bearing failure in rotating equipment, such as motors, pumps, and compressors, on a manufacturing line, using vibration, temperature, and load data. Its declared operating domain is the specified equipment classes, the sensor suite, and the stated operating ranges. In the national risk tiering of Section 3, such a system operating with human oversight falls in the medium tier.

In the laboratory (Band 1), the system is first specified with explicit reliability targets, for example a required detection lead time and a bound on false alarms, together with a hazard analysis (APRL 1). It is then validated on held-out historical data spanning the declared equipment and operating ranges (APRL 2). Finally, it is stress-tested against sensor dropout, previously unseen fault signatures, and load and temperature excursions; its failure modes are catalogued; and a safe fallback, reversion to scheduled maintenance, is specified (APRL 3).

In the pilot (Band 2), the system is deployed on a supervised line where maintenance engineers retain oversight and can override its recommendations, with live monitoring of precision, recall, and lead time (APRL 4). It sustains these metrics across several months and varying production conditions, with drift detection active and a recovery exercised after a missed fault (APRL 5). It reaches controlled release once residual risk is reduced to ALARP and its assurance case is independently verified (APRL 6).

In operation (Band 3), the system is proven across the plant's full population of covered equipment over a defined period and entered in the national registry (APRL 7), and it then maintains performance through seasonal load changes and rare fault types, without safety-critical failure, over an extended period (APRL 8).

Suppose the plant later introduces a new class of high-speed pump that lies outside the declared operating domain. Monitored recall on the new equipment falls below the sector threshold, and drift detection flags the change. Under the revocability rule, the system's rating for the affected equipment is suspended and revised from APRL 8 to APRL 5 pending re-validation on the new class, while its rating for the original equipment is unaffected and its operation continues. The example illustrates the two properties that distinguish APRL from a one-time certificate: reliability is always measured within a declared domain, and a rating is maintained against continuing evidence rather than granted once. The specific metrics and thresholds used here are illustrative and would be fixed by the relevant sector code.

Vienna International Centre
Wagramerstr. 5, P.O. Box 300,
A-1400 Vienna, Austria

+43 1 26026-0

www.unido.org

unido@unido.org

UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION