Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Tel 425 882 8080
Fax 425 706 7329
www.microsoft.com

Learning from other Domains to Advance AI Evaluation and Testing

We are grateful to the authors of the enclosed expert report, which forms part of a broader series commissioned by Microsoft.

These reports were commissioned as part of Microsoft's effort to draw lessons from other domains to strengthen testing and evaluation as a cornerstone of AI governance.

The insights contained in this report reflect the authors' independent analysis and expertise. The views expressed are those of the authors alone.

We thank the authors for their intellectual generosity and thoughtful engagement throughout this project.

Full Series

Civil aviation: *Testing in Aircraft Design and Manufacturing* by Paul Alp.

Cybersecurity: *Cybersecurity Standards and Testing—Lessons for AI Safety and Security*, by Stewart Baker.

Financial services (bank stress testing): *The Evolving Use of Bank Stress Tests*, by Kathryn Judge.

Genome editing: *Governance of Genome Editing in Human Therapeutics and Agricultural Applications*, by Alta Charo and Andy Greenfield.

Medical devices: *Medical Device Testing: Regulatory Requirements, Evolution and Lessons for AI Governance*, by Mateo Aboy and Timo Minssen.

Nanoscience: *The regulatory landscape of nanoscience and nanotechnology, and applications to future AI regulation*, by Jennifer Dionne.

Nuclear energy: *Testing in the Nuclear Industry*, by Pablo Cantero and Gerónimo Poletto Antonacci.

Pharmaceuticals: *The History and Evolution of Testing in Pharmaceutical Regulation*, by Daniel Benamouzig and Daniel Carpenter.

Testing in the Nuclear Industry

Authors: Pablo Cantero[1] – Gerónimo Poletto Antonacci[2]

Contributors: Pablo Ramirez[1], Augusto Debandi[1] and Federico Mezio[1]

1. Instituto Balseiro, UNCuyo – CNEA 2. Autoridad Regulatoria Nuclear Argentina

1. Introduction

The growth in global energy demand accelerated in recent years (2023–2024) at rates double the average growth of the past decade (2010-2019)[1]. Projections suggest this trend will continue. Combined with the global goal of decarbonization, this positions nuclear energy as a reliable source with significant potential for growth.

However, the benefits of the nuclear industry go beyond electricity generation and extend to healthcare, scientific and technological research and development, food security, and environmental protection among many other activities.

The nuclear industry is technologically mature and constantly evolving. The main associated risks are related to potential accidents and their consequences, including environmental contamination and radiological harms. Alongside the evolution of the nuclear industry, a consistent regulatory framework has been developed, evolving over time to keep up with new technologies and the operational experience of nuclear facilities. Different regulatory approaches exist in different jurisdictions. While some establish prescriptive regulations (i.e. defining how to do things), others use performance-based regulations (i.e. establishing a limit of risk and leaving the method of compliance to the designer). These approaches have in common that they establish standards aiming to ensure, with a high degree of certainty, that activities are conducted safely. In pursuit of this objective, testing provides an objective means of demonstrating compliance with specific requirements stated in regulatory or industry standards. However, testing is not only performed to ensure safety and regulatory compliance, it also supports and is an essential tool in maintaining the capability and availability of nuclear facilities to provide intended goods and services such as production of electricity, medical radioisotopes, neutron beams, etc.

Testing is not only the specific action of evaluating a structure, system or component (SSC) to ensure that it functions as it should. It encompasses the whole process of evaluating specific characteristics of SSCs or even a facility in its entirety, by comparing them against established acceptance criteria. These criteria may refer to design objectives established by the design authority, regulatory requirements according to applicable regulation, or international industrial standards. Therefore, the testing process is not limited to conducting trials but also takes advantage of various systematic resources, which may include analyses (based on models and calculations) or evaluations (based on a specific implementation), inspections, application of experience (drawing from a history of successful operation), similarity (i.e. comparing with similar solution that works

[1] Global Energy Review 2025 - International Energy Agency – https://www.iea.org/reports/global-energy-review-2025

properly and demonstrating, given fundamental similarities, that the system or component will behave the same), among other options.

2. Overview of the testing landscape in nuclear

Testing in the nuclear industry is present during the entire lifecycle of a facility, from the conception, throughout design, construction, commissioning and operation stages. Licensing accompanies all these stages as a comprehensive process. Even the final decommissioning of a nuclear facility is subjected to testing. Even though nuclear fuel has already been removed and there are no longer operational concerns, testing assists the decommissioning process with a focus on sustaining the radiological safety of people and the environment long after a facility has ceased operation.

Testing is primarily driven by two aims that guide the definition, configuration, and execution of tests: **safety of the facility**—i.e., operating the plant while minimizing potential risks of a nuclear accident—and the **capability and availability of the facility to provide intended good and services**. In both cases, the goal is to objectively demonstrate that systems meet the defined acceptance criteria for achieving their stated objectives.

For safety testing, Argentina, a jurisdiction that adopts a performance-based regulatory approach, defines risk as the product of the probability of occurrence of an accident sequence, and the health consequences for a critical group of persons that might be generated by that accident sequence. The risk for each potential accident sequence is evaluated by performing safety assessments and then compared to the acceptance criteria established in regulatory standards.

Each stage of the life of a facility has a specific scheme of testing that is shaped by the objective pursued. Table 1 provides an overview of testing in the nuclear industry lifecycle.

2.1 Testing in Design and Construction

During the design stage, testing is mainly used for:

- the validation of simulation and modeling codes; and
- the verification of the final design of SSCs that:
    - include innovations; or
    - are intended to perform under harsh environmental conditions; or
    - fulfill functions for which there is no precedent.

For validation of simulation and modeling software, these are tested against models whose results are known. This process, known as benchmarking, is inherited from the scientific community, which generally also develops the software and agrees on which testbenches[2] should be run to establish their proper functioning for the intended application, thus validating its utilization by industry.

[2] Controlled environment used to demonstrate or validate the operation of a system or device.

For verification of the design of SSCs, the testing process is carried out on a prototype SSC built to accurately represent the characteristics that are intended to be validated. The test conditions must be representative of anticipated operating conditions for the tested equipment. The test results are compared with acceptance criteria, which are derived from the design basis[3] outlined in the design of the SSC under test.

Beyond assessing functional performance, the time the equipment remains operational is also important, so there is a need during the equipment design phase to consolidate materials selection or components design. For these purposes, the testing scheme is similar but emphasizes materials used in the prototype and environmental conditions of the test, which must be consistent with, or even exceed normal operating conditions, in order to extrapolate the wear and tear that could affect the entire lifespan of the equipment.

2.2 Testing in Commissioning and Operation

Commissioning and operation of a nuclear facility is directly supported by testing to demonstrate the safety and the capacity of the plant. However, the approach to testing differs from that taken during the design and construction stage. While earlier testing ensured SSCs meet design criteria, during successive stages, testing is used to verify safe and efficient operation under conditions defined by the design authority in accordance with applicable regulation and production objectives.

During these stages, the design authority is responsible for establishing the procedures and acceptance criteria for testing against the technical specifications and criteria authorized in the license (see section 2.3), typically guided by international standards and known good practices (see section 4). The test results must be reviewed and approved by the regulatory authority to ensure compliance with established requirements.

Testing is critical during commissioning. It involves comprehensive tests to verify the facility follows design and regulatory requirements, which aim to ensure it can sustain safe operational conditions. This process typically includes three or more sub-stages: (1) pre-operational testing, (2) loading of radioactive or fissile materials and (3) initial criticality and staggered power increments testing.

Pre-operational testing seeks to verify SSCs conform to the design basis and meet performance criteria. A primary objective in this phase is to validate the proper functioning of safety systems, essential to safely load of nuclear material. Once nuclear fuel is loaded, testing is central to validating the nuclear configuration, thus aligning the actual results with design parameters and analyses. Upon fuel loading, testing starts under safe-mode conditions (e.g., initial criticality, low power operation) and progressively expands to nominal conditions, verifying the facility operates within safety limits and confirming the estimates and hypotheses related to the facility's behavior.

During operation, testing continues in critical areas such as safety, availability, plant optimization, environmental impact and radiation protection.

---

[3] Design basis refers to the set of conditions for which a specific design in performed.

Regular periodic testing verifies the functionality and reliability of safety systems, ensuring they meet their defined technical specifications. Furthermore, aging surveillance programs continuously monitor material degradation to maintain the reliability of SSCs over time.

Testing also plays a crucial role in maintaining operational efficiency by minimizing downtime and unplanned outages. This is achieved through predictive and preventive maintenance programs that extend component lifespans, and through testing of SSCs, which act as a proxy for the facility's general condition and inform the enhancement of operational outcomes.

The environmental impact monitoring of a nuclear facility relies on testing to assess radioactive effluent releases and the proper storage and disposal of radioactive waste and spent fuel, ensuring compliance with regulatory limits.

Equally important to all previous considerations is radiation protection. This is inherently important to all aspects of a nuclear facility's operation. Any activity involving radiation must prioritize workers' and public safety and must adhere to strict regulatory dose limits. In this context, testing is indispensable for dose management, aiming to minimize personnel exposure while ensuring adequate workforce availability for tasks that directly influence the facility's performance.

Once the facility has reached the end of its operational life, testing is a cornerstone of safety considerations during the decommissioning process. The SSCs to be dismantled are categorized based on different types of measurements and tests. This classification supports the selection of suitable methods for managing and disposing of potentially hazardous nuclear waste.

Table 1. Overview of testing across a nuclear facility lifecycle

Lifecycle stage	**Purpose of testing**	***Id test***
Design and Construction	Validation of simulation and modeling codes	Benchmarking
	Verification of final SSCs design	Testing on prototypes
	Durability assessment	Endurance testing
Commissioning	Verification SSCs conform to design basis	Pre-operational testing
	Validation of safety systems	Testing of reactivity-regulating, cooling, confinement, I&C[4], and support systems
	Validation of nuclear configuration	Measurement of nuclear parameters
	Verification of safe operation	SSCs capability tests under initial criticality, low-power and nominal conditions.
Operation	Verification of safety systems	Periodic testing

[4] I&C stands for 'Instrumentation and Control' and is used to refer to I&C equipment or I&C functions.

	reliability and functionality	
	Aging surveillance	Material degradation testing
	Operational efficiency	Testing for SSCs maintenance
	Environmental impact monitoring	Monitoring of radioactive effluent releases
	Radiation protection	Testing for radiation dose management
Decommissioning	Waste management	Testing for waste categorization

2.3 Testing for Licensing

Nuclear facilities and activities are required to comply with regulations related to safety, safeguards, and physical security to ensure the protection of the public, workers and the environment. Typically, each country establishes an organization that serves as the nuclear regulatory body. For example, this role is played by the Nuclear Regulatory Authority in Argentina; the Nuclear Regulatory Commission in the United States; and the Authority for Nuclear Safety and Radiation Protection in France.

Any organization seeking to operate a nuclear facility is required to apply for licenses from the regulatory body for its siting, construction, commissioning, operation and decommissioning.

Different countries adopt different regulatory approaches. Some countries opt for prescriptive regulations while others favor performance-based regulations. A regulatory framework based on prescriptive regulations may go as deep as establishing precise procedures for testing. On the other hand, in a performance-based framework, the licensee organization proposes specific criteria—based on recognized standards—to demonstrate compliance with regulatory requirements, which must be agreed upon by the regulatory body.

For instance, the Argentinean regulatory requirements are performance-based while in the United States and France the regulatory bodies establish requirements that are more prescriptive. In all cases, regulatory requirements address both normal operation and accident conditions. Basic regulatory requirements usually set limits on the radiation dose that workers and the public can be exposed to, while other more specific requirements apply to safety-critical SSCs, e.g., reactivity-regulating, cooling and confinement systems, building and structures, I&C and support systems. For example, under normal operating conditions of a nuclear installation, a radiation dose limit of 20 units[5] averaged over five years is applied for workers with no single year exceeding 50 units. For the public, a radiation dose limit of 1 unit per year is typically enforced.

[5] The unit referred in the example is mSv. The sievert (Sv) is a unit of measurement for radiation dose. It quantifies the amount of radiation absorbed by a person.

The performance-based approach allows the licensee organization to propose specific criteria based on recognized standards to demonstrate compliance with regulatory requirements.

Test results are used directly to verify compliance with the adopted acceptance criteria or as an input into other methods for demonstrating compliance such as analysis of normal operating or accident conditions, use of operating experience, or a combination of these approaches. The licensee develops testing plans, which must address all the specific criteria to be verified. Once the plan is approved by the regulatory body, the licensee organization is typically in charge of conducting testing while the regulatory body oversees it and reviews the testing results. Additionally, during any licensing stage, if a major modification to the design of a SSC is necessary, regulations require the licensee to demonstrate that the modification does not deviate from the requirements imposed by the license. The licensee may be required to conduct testing to qualify the modified SSCs and demonstrate compliance with the acceptance criteria established in the license. Testing may even require that the affected part of the facility undergo commissioning tests to verify compliance.

3. The history of testing in the nuclear domain

The civil nuclear industry emerged as a spin-off from military research. Initially, countries leading in nuclear technologies established organizations for the promotion and development of the nuclear field. This was true of the U.S. Atomic Energy Commission and, in Argentina, the National Commission of Atomic Energy (CNEA).

During this early phase, testing primarily served to validate theoretical designs of nuclear reactors, materials, technologies, systems, etc. While regulatory oversight was limited at the time, some form of control was established, mainly through a subsidiary division of the promotional organization itself.

Over time, nuclear designs were consolidated in different countries, and regulatory frameworks also evolved, often featuring requirements specific to domestic technology designs. Different approaches to regulation were also consolidated. Some countries opted for prescriptive regulations while others favored performance-based regulations. This gave rise to different approaches for testing compliance with those regulations. A regulatory framework based on prescriptive regulations may go as far as establishing detailed procedures for testing. Testing within a performance-based framework instead focuses on robust demonstrations of compliance and the exact testing methodologies to be used are negotiable with the regulator.

Increasing criticism related to the close ties between organizations primarily focused on promoting nuclear technologies and regulatory bodies led to the formation of international consensus that countries using nuclear energy should strive to institute an independent regulatory body that would be exclusively concerned with safety. The formation of this consensus was a gradual process, driven by growing concerns about the potential risks of nuclear energy and the need for robust and independent regulatory oversight. The Convention on Nuclear Safety, described below, played a crucial role in achieving it.

The Argentinian case exemplifies this. Initially, oversight functions were integrated within the National Commission of Atomic Energy. During the 1990s, the oversight unit was separated from the Commission and became the autonomous Argentinian Nuclear Regulatory Authority (ARN). Regulations were issued by ARN covering all stages of the nuclear facility lifecycle, forming the basis for design, construction and operational testing. The performance-based framework adopted enabled flexibility in testing different designs against requirements.

The occurrence of key disasters reshaped regulatory frameworks worldwide. The Three Mile Island accident in 1979 made clear the need for tighter control. Staff training, the handling of abnormal situations and the consideration of human factors led to regulations strengthening their focus on quality management of operational processes.

After Three Mile Island, several measures were taken to further reduce the risks of nuclear operation through engineering measures, better management practices, and more competent operation. These responses were widely and openly adopted by Western actors. The opaque political landscape at the time impeded an open discussion about whether similar risk analyses and measures were undertaken by Soviet reactors.

Chernobyl in 1986 underscored the dangers of isolation and self-reliance of nuclear industry players. The event led to a realization that an accident of anyone is an accident of everyone because the consequences of Chernobyl hindered the nuclear industry worldwide for decades. Experts' post-accident analyses of the disaster was only achieved after intense international pressure for Soviet agreement.

Chernobyl, and to a lesser extent Three Mile Island, dramatically highlighted the transboundary consequences of a nuclear accident, spurring the International Atomic Energy Agency (IAEA) and its members to convene an international conference on the safety of nuclear power in 1991. The main outcome of this conference was the initiation of an international Convention on Nuclear Safety under which the contracting states would voluntarily commit themselves to a series of requirements to strengthen domestic nuclear safety and international cooperation. The Convention was adopted in 1994 and entered into force in 1996.

On its own volition, the nuclear industry sought the establishment of an international safety net to support and enhance safety through peer-reviewing each other's practices and sharing best practices among members. Commitments of this sort had been previously attempted, but they mainly gathered smaller, intra-national groups of nuclear operators. The objective this time was to engage the international industry. Thus, the World Association of Nuclear Operators (WANO) was created, bringing together most nuclear utilities worldwide.

This altered the testing landscape: in addition to meeting regulatory requirements, operators now face the scrutiny of peer reviews. This entails demonstrating compliance with industry best practices, often tested through benchmarking and peer comparisons.

Fukushima in 2011 forced a re-examination of assumptions used for testing. Before the accident, focus was centered on design and testing for high likelihood scenarios. After the accident, a worldwide review of plant designs, known as "stress-testing", was set in motion. The purpose was to test the robustness of designs against accidents of very low probability that would push the systems beyond their design basis. Thus, an extended set of design and testing requirements were developed, triggering a wave of design improvements and retrofitting of nuclear plants.

4. International coherence and interoperability

International coherence within the nuclear industry is primarily driven by multilateral organizations such as the International Atomic Energy Agency (IAEA) and the Nuclear Energy Agency (NEA). These institutions collect and share—in the form of standards, guides, technical documents, etc.—the international consensus regarding a wide range of issues.

Related to testing, the IAEA develops standards within its Safety Standards publications. These standards cover diverse stages in the lifecycle of a facility such as construction, manufacturing, commissioning and decommissioning. They indicate what should be covered by testing during each stage, such as nuclear installation safety testing, non-destructive testing, pressure testing, functional testing, and radiation protection testing, etc. Additionally, the IAEA encourages the sharing of experience internationally, reflected in other publications wherein users can find the latest scientific knowledge, best practices and specific examples on how testing is conducted in other countries.

Standards published by the IAEA are not legally binding for member states and are seldom adopted by national regulators as their actual regulations. However, they serve as a reference point for the minimum requirements that a sound nuclear design or regulatory framework should address. This guidance is especially valuable for newcomer countries with no prior nuclear-industry experience that aim to integrate nuclear technology into their national infrastructure.

By contrast, countries with mature nuclear industries have developed their own regulations and industry standards for evaluating and licensing facilities. As technology leaders, these countries are typically also exporters of nuclear technologies. As a result, the cross-border deployment of nuclear designs is generally made between countries that share the same technological lineage. Therefore, technical standards and requirements, initially established for domestic designs, have been often transferred to countries importing nuclear technologies. But those same requirements if applied to a foreign design could render it non-authorizable.

Despite this heterogeneity, fundamental safety and overarching core testing requirements are well agreed and shared among the international community. The evolution of this shared understanding is strongly shaped by history, tracing back to the foundational principles of the early nuclear-leading nations. Rather than independently developing their own standards, countries subsequently entering the nuclear industry adopted these established principles, effectively elevating them to de facto standards. Organizations such as IAEA, NEA and WANO further formalized this convergence by incorporating their underlying philosophies into fundamental safety requirements.

Looking ahead, the nuclear industry is making considerable efforts to harmonize testing approaches in anticipation of growing interest in Small Modular Reactors (SMRs). Imposing new, location-specific testing requirements at each site would significantly slow the deployment of these facilities across diverse economic, geographic, and regulatory settings. Harmonization is still a distant goal. The SMR sector remains in its infancy, experimenting with numerous—and sometimes unconventional—designs, so shared testing standards must wait until many foundational definitions are settled. The difference of current developments, compared to historical ones, is the recognition for the need to define testing requirements almost in parallel with the evolution of designs, rather than as an afterthought as was often the case in the past.

5. Domain lessons learned and recommendations

The nuclear industry is grounded in proven technologies and upholds some of the world's highest reliability standards. This record reflects decades of continual refinement—of analytical methods, operating procedures, quality-assurance programs, and technical norms—all focused on improving reactor performance and further reducing the risks associated with nuclear facilities and power plants. As noted above, the industry's strides toward greater safety and efficiency have benefited from the diverse approaches adopted by its stakeholders. One key lesson learned is that dispersion of designs or plant technologies increases inefficiency of operation of a nuclear fleet, as it requires the development of specific testing activities particular to the features and acceptance criteria of each facility. Testing results are also limited in their relevance, as they only apply to a particular design. On the contrary, harmonization makes operation more efficient and safer. As testing requirements and procedures are harmonized between several facilities and organizations, common efforts undertaken to streamline testing mechanisms translate into widespread efficiencies. Those gains boost overall performance and, because safety improvements now spread quickly throughout the industry, further enhance safety.

The nuclear industry has also benefited significantly from the establishment of multilateral institutions. They facilitated the formation of consensus on key nuclear issues, enabled countries to join forces in solving common problems and fostered the sharing of experience and best practices among stakeholders in the nuclear industry. The imperative for global governance is also driven by the recognition that the actions (and failures) of any individual organization have a global impact. By setting up international bodies that function as quasi-watchdogs, the global community spurred a concerted push to raise minimum safety-testing standards worldwide.

Looking ahead, artificial intelligence is set to reshape testing in the nuclear industry. Early uses—such as faster, smarter querying of vast operating-experience databases—are already appearing. Even more advanced applications are under active discussion: expert-system decision support, automated testing and inspection, and digital-twin simulations. Once validated for nuclear contexts, these technologies could profoundly transform the sector's testing landscape.