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 author 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 each 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 Aircraft Design and Manufacturing

Paul Alp
Adams and Reese LLP
Washington, D.C.

January 2025

I. Introduction

In aviation design and manufacturing, testing can generally be described as serving the following purposes: (i) facilitating development of a product design; (ii) verifying its functionality and robustness; (iii) demonstrating compliance of the design with regulatory standards in order to obtain approval from a civil aviation authority such as the U.S. Federal Aviation Administration; and (iv) verifying that the as-built product conforms to its design, complies with applicable regulations, and is in a condition for safe operation.

Testing is, therefore, closely linked to showing compliance with regulatory requirements, which forms the predicate for obtaining approvals from aviation authorities that allow aircraft to be operated. To be sure, some design features do not require testing to show compliance, and compliance showings may also be made based on analysis, historical findings of compliance, inspection, or system architecture. But, testing remains a centerpiece of compliance, and the regulations to which compliance must be shown fundamentally define the nature of aircraft products and the processes for creating them.

For these reasons, testing with respect to aviation products must be considered in the context of the overarching regulatory framework for standards and how compliance to standards must be shown.

II. Overview of the Testing Landscape in Design and Manufacturing

The international aviation regulatory regime is comprehensive and mature. Civil aviation authorities like the FAA impose detailed requirements on the design, manufacture, and operation of aircraft. The design of an aircraft and its component parts must satisfy regulatory requirements, known as "airworthiness standards," in order to be eligible for approval by regulators. Testing is fundamental to demonstrating compliance with airworthiness standards. After a design is approved, a product must be manufactured in accordance with the design, and testing is used to demonstrate conformity of the as-built product to the design.

A person or entity seeking approval from an aviation regulator is typically referred to as an "applicant." An applicant must make "showings" of compliance with regulatory requirements. A regulator like the FAA issues approval such as through a type certificate when it makes "findings" of compliance. Regardless of whether applicants perform testing in connection with design or manufacturing, they must adhere to strict requirements regarding traceability: the

** This document is provided solely for information purposes and is intended to provide information to assist the reader in implementing its own analysis and review of the topics discussed. Nothing contained herein constitutes legal advice.*

creation and maintenance of a paper trail sufficient to document results of tests and inspections used to show compliance in support of certification.

A. Testing in Design

Testing is integral to aircraft design because it forms the basis for obtaining regulatory authorization. For aircraft, such authorization is typically in the form of a "type certificate." The FAA issues a type certificate when it finds that the design of an aircraft complies with applicable airworthiness standards. An applicant that designs a new aircraft or major modification to an aircraft typically uses testing extensively to show that the design complies with airworthiness standards in order to obtain a type certificate.

Airworthiness standards are issued by civil aviation authorities of individual countries. Pursuant to the Chicago Convention, discussed below, the International Civil Aviation Organization ("ICAO") promulgates standards that signatory countries are encouraged to follow under a framework that supports reciprocal recognition of authorizations based on harmonized standards. Accordingly, regulators generally seek to harmonize their airworthiness standards. Indeed, standards issued by ICAO are commonly driven by major aviation authorities such as the FAA and its European counterpart.

Airworthiness standards vary from broadly worded to highly specific, and apply to almost every aspect of an aircraft, as shown by the following examples from FAA requirements:

- "Each flight, navigation, and powerplant instrument for use by any pilot must be plainly visible to him from his station with the minimum practicable deviation from his normal position and line of vision when he is looking forward along the flight path."

- "The airplane must be shown by analysis, test, or both, to be capable of continued safe flight and landing after any [specified] failures or jams in the flight control system within the normal flight envelope."

- "Each electrical and electronic system . . . must be designed and installed so the system is not adversely affected when the equipment providing these functions is exposed to equipment [High Intensity Radiated Field] test level 1 or 2. . . ."

Companies developing products that will ultimately require certification typically move as quickly as possible from functional development into testing against regulatory criteria in order to ensure that their designs satisfy airworthiness standards.

The context in which testing takes place can include both a specific technology or general technology applied in the specific context. An example of a specific technology would be a collision avoidance system through which computers in individual airplanes communicate with each other to identify situations where a mid-air collision may take place, and then recommend evasive action to the respective pilots of the conflicting airplanes. The system, known as a traffic alert and collision avoidance system ("TCAS") was originally developed by the FAA in collaboration with industry, mandated by FAA in its regulations, and subsequently incorporated in ICAO standards applicable to airliners worldwide.

General technologies imported into the aviation domain must also satisfy airworthiness standards. For example, video displays have non-aviation uses, but, in the cockpit of an airplane, they may be required to satisfy requirements relating to matters such as reliability and human factors concerns such as visibility from different viewing angles.

Correct thresholds and measurement criteria may be decided either (i) through prescriptive requirements contained in regulations or (ii) by industry participants, typically working cooperatively through standards organizations that develop consensus means for demonstrating compliance with airworthiness standards. Over time, regulators have increasingly accepted standards and means of compliance developed by industry. Although no clear bright line can be stated for when a regulator will choose one approach over the other, regulators have generally been more accepting of industry-developed standards where safety concerns are less critical. For example, a small homebuilt aircraft that carries two occupants presents less safety risk than an airliner that carries 400 passengers, and – due to the magnitude of risk presented – in the former case regulators are more open to industry standards, while in the latter case the regulators may impose detailed requirements. But, as discussed below, even where prescriptive detailed regulatory requirements exist, regulators commonly accept methods of compliance developed by industry.

An example of highly prescriptive regulatory requirements would be where a regulation states that a particular part must satisfy a specific standard of performance defined by objective criteria set forth in regulations which also prescribe how testing to demonstrate compliance must be performed. For example, materials used in the cabin of an airliner must satisfy various flammability standards, and these standards state that, in order to be compliant, a sample of the material must be shown to meet specific performance requirements when tested using methods, tools, and acceptance criteria that the regulations describe in detail. FAA regulations contain almost 14,000 words describing criteria and test methods that must be used for such flammability tests, reaching matters such as the precise size and configuration of specimens, the apparatus to be used, test conditions, and criteria for acceptance such as:

> For at least two-thirds of the total number of specimen sets tested, the burn length from the burner must not reach the side of the cushion opposite the burner. The burn length must not exceed 17 inches. . . . .
>
> The average percentage weight loss must not exceed 10 percent. Also, at least two-thirds of the total number of specimen sets tested must not exceed 10 percent weight loss. . . . .

(Source: Appendix F to 14 C.F.R. Part 25, "Test Criteria and Procedures for Showing Compliance with § 25.853 or § 25.855").

Other airworthiness standards do not prescribe testing requirements in such detail, or at all. How these standards will be satisfied and how compliance will be shown may be worked out on a case-by case basis in connection with a specific project where the applicant reaches agreement with the regulator. For example, a regulation may state that an aircraft must be capable of controlled flight when an engine fails. The aircraft designer may develop sophisticated means for ensuring that the aircraft remains in control, and then will show, through methods agreed by the regulator, how such systems achieve the desired end-state of maintaining aircraft control.

It is, therefore, common for a back-and-forth to take place between applicant and regulator over what acceptable methods of showing compliance are, although it is the FAA that ultimately decides. Methods of showing compliance could include the system design itself, testing, inspection, quantitative and qualitative analysis, showing similarity to previously approved products, conformance to industry consensus standards, or demonstration. The FAA has a formal, structured process to work with an applicant to determine the appropriate method of compliance. In this "issue paper" process, (i) the FAA creates a living document called an "issue paper" in which it states its position on a certification issue; (ii) the applicant adds its responsive position to the paper; (iii) the parties confer and update the paper with new information as necessary; (iv) when they align on the method of compliance it is memorialized in the paper. The issue paper formally documents how compliance with a specific requirement will be shown, and it becomes a definitive record of how the issue was resolved. An applicant may rely on an approved issue paper in subsequent certification projects where similar compliance issues are presented.

In many instances, industry participants work collectively through standards organizations such as SAE, International (formally the Society of Automotive Engineers) and the Radio Technical Commission for Aeronautics, Inc. ("RTCA") to develop consensus standards and guidance on processes, criteria, and testing methods used to satisfy regulatory requirements. The FAA and other regulators often participate in meetings of these organizations and issue formal guidance documents that accept methods of showing compliance developed by industry organizations. For example, the performance standards for TCAS equipment described above were defined in technical documents developed by a joint regulator-industry working group and published by RTCA.

Testing takes place in phases. The initial phase is marked by functional development and product iteration, followed by a verification phase in which the applicant assesses the functionality and robustness of the product. Each of these phases is part of the applicant's in-house development process. Once the applicant is satisfied with the product, testing moves into a showing phase in which the applicant demonstrates compliance to the regulator. How, as a practical matter, testing fits in to an airplane type certification project with the FAA can be described as follows.

The path to certification begins with meetings between the applicant and the FAA about the nature and scope of the project. These discussions include identifying the appropriate "certification basis," which is an itemization of the specific regulatory requirements (airworthiness standards) to which the applicant will show compliance. The certification basis is ultimately established by the regulator and agreed by the applicant, based on a mutual understanding of the product's design features. The applicant develops a certification plan, which discusses the certification basis and describes how compliance will be shown with each requirement of the certification basis, such as through testing, analysis, inspection, or demonstration as appropriate. The plan is finalized and approved before the project moves into the phase where the applicant shows compliance to the regulations.

During the product development phase, various aspects of the certification plan and certification basis may change. When the design has sufficiently evolved, the applicant performs testing and analysis to develop data to be used for showing compliance. Testing may be performed by the applicant itself or by specialized subcontractors. Eventually, after the certification plan is finalized and approved, the applicant provides the FAA with a package of reports, test results, and substantiating materials to support its claim that the design is compliant. The FAA reviews the applicant's submission and, if it is acceptable, initiates formal FAA inspections and tests, including ground and flight testing as the Agency deems necessary, to assess the final design for compliance with regulations.

At the compliance showing and finding phase, the FAA, within its discretion, may accept test results and data presented by the applicant, witness tests performed by the applicant, or perform tests and analysis itself. All such methods are commonly used. The FAA determines what method may be appropriate in a given instance, and the method may be driven by the Agency's staffing needs and not inherent aspects of the design feature to be assessed.

Once all inspections and tests required by the FAA are completed and the Agency finds that the technical data for the project complies with applicable regulations and the airplane conforms with its design and can be operated safely, the FAA will formally approve the design by issuing a type certificate.

In sum:

- Applicants perform testing to design and validate their products and also to show compliance with applicable airworthiness standards.
- The airworthiness standards are set forth in regulations issued by national civil aviation authorities that are harmonized internationally to a large extent.
- Airworthiness standards can be:
  - Highly prescriptive, setting forth design and performance criteria as well as methods for testing and showing compliance;
  - Prescriptive but not inclusive of methods for testing and showing compliance, which may be established by agreement between the applicant and regulator on a case-by-case basis, through reference to industry-developed standards, or both;
  - Largely based on industry consensus standards that are developed in coordination with regulators.
- The project certification plan functions as a road map that memorializes applicable airworthiness standards and methods that will be used to show compliance to each.
- Compliance may be shown through design, analysis, testing, inspection, or a combination of methods.
- In making compliance findings, the FAA has discretion to rely on data provided by the applicant, observe tests performed by the applicant, or perform tests and inspections itself.

B. Testing in Production

Regulated entities obtain authorization to manufacture products that conform with approved designs. An aircraft manufacturer's essential concern is to ensure that the as-built product is conforming. Accordingly, during production and after the product is completed, testing is used to assess functionality as well as to verify that the product conforms to its design and satisfies regulatory requirements.

When the applicant believes the product is complete and conforming, it formally presents the product to the FAA for review. At this stage, the FAA will perform inspections or tests that it deems necessary to make conformity findings, an example being functional flight testing to assess a completed aircraft. The FAA has full discretion to determine the appropriate means for performing these conformity activities. It may, for example, perform inspections or tests itself, observe inspections or tests performed by the applicant, allow designees with authority conferred by the FAA to perform inspections and tests on the FAA's behalf, or accept data provided by the applicant. In complex projects, the FAA commonly uses all of these methods. When the FAA determines that a newly manufactured aircraft conforms to its approved design and is in a condition for safe operation, it will issue an "airworthiness certificate" that memorializes the FAA's approval to put the aircraft into operation.

III. History of Testing

Early aviation pioneers worked in a largely unregulated domain and used functional testing to iterate on their ideas. Although governments took intense interest in activities of aviation pioneers, they were reluctant to issue regulations that would potentially stifle innovation. The thinking at the time was that detailed regulations would slow or discourage development of new technology in the public interest. Accordingly, governments allowed inventors to tinker, experiment, and fail in furtherance of the long-term goal of innovation. Congress passed the Air Commerce Act of 1926 in part due to the urging of industry for federal action that would support development of air transportation. The Act's stated purpose was to "encourage and regulate the use of aircraft in commerce, and other purposes." The Act conferred on the Secretary of Commerce the "duty" to foster air commerce, while also requiring promulgation of basic air traffic rules that would allow safe operations and a scheme for issuing certificates to pilots.

As time passed and serious incidents and accidents took place, governments issued increasingly prescriptive regulations. A common truism is that most aviation regulations are written in blood. As the need for safe and repeatable operations exceeded the need for innovation for its own sake, the prevailing approach to regulation evolved into a framework of detailed requirements to which compliance must be shown through various means, particularly testing and inspection.

As aviation matured into the 21st Century, approaches to standards-setting and compliance-showing evolved. Instead of defaulting to detailed regulations which specify that articles must have specific characteristics, as shown by specific kinds of testing and analysis, regulators moved to a more performance-based approach for many regulations governing aircraft design. Under a performance-based paradigm, regulations set forth general functional requirements at a high level and industry participants develop specific means of complying with, and demonstrating compliance to, such requirements. Frequently, detailed criteria, means of showing compliance, and test methods are developed through the consensus of industry participants working through standards organizations. The performance-based approach is intended to encourage innovation, foster efficiency, and ensure that design standards keep pace with rapidly evolving technologies. This approach is generally perceived as successful and is becoming increasingly common, both in replacing old airworthiness standards for legacy product types and developing new standards for novel technologies such as drones and advanced air mobility aircraft (*e.g.,* quadcopters used as air taxis).

Interestingly, consistent with the early period of aircraft development, regulators today have taken a relatively hands-off approach to regulation of commercial space launch activities.

Regulatory agencies have intentionally held off on issuing detailed prescriptive regulations where possible in order to support innovation in this new field of endeavor. For example, the FAA's regulatory framework governing commercial space launch activities largely centers on the FAA performing safety reviews of vehicles and ground facilities to support the issuance of launch licenses. It does not involve design, performance, and testing requirements like those for aircraft that are set forth in regulations. As commercial space launch activities mature, the regulatory framework is expected to evolve toward increasing prescriptive requirements.

IV. International Coherence and Interoperability

The international aviation regulatory framework has developed to be coherent and harmonious with respect to certification standards. Testing standards, methods, and criteria are commonly consistent and interoperable across international boundaries. An international regime based on treaties forms the cornerstone of this domain.

The 1944 Convention on International Civil Aviation, known as the "Chicago Convention," established the current international aviation regulatory framework. It describes the roles and authorities of the contracting states in a variety of matters, and is a living document with annexes setting forth standards and recommended practices with respect to design, certification, and operation of aircraft. It generally provides that contracting states will, to the extent consistent with their own laws, recognize authorizations issued by other states where the criteria on which the authorizations are based are either consistent with, or stricter than, the criteria for such authorizations set forth in the Convention. Its annexes are continually revised and updated by agreement of contracting states, and ICAO, an agency of the United Nations, is responsible for its development and administration.

In addition to the Chicago Convention, many countries have executed bilateral agreements and corresponding technical implementation procedures that set forth in detail how reciprocity and recognition of authorizations under the Convention shall work, and how regulators cooperate and coordinate to harmonize standards and practices. For example, the country with jurisdiction over an aircraft designer is responsible for approving the design of the aircraft by issuing the original type certificate. When an aircraft with a type design approved by one state is exported to another, the importing state may accept the aircraft if it satisfies the importing state's airworthiness standards, under a process known as "validation." The validation process is intended to rely on the type certificate issued by the state of design to the extent possible. The state into which the aircraft is imported will, therefore, generally accept showings of compliance and results of tests performed in connection with original design approval provided that the two countries' airworthiness standards and compliance processes are consistent and compatible.

Because countries attempt to harmonize their airworthiness standards and acceptable means of compliance as much as possible, acceptable testing methods are also frequently consistent and harmonized across international boundaries. This consistency and coherence is enhanced through the use of industry organizations to develop standards, testing methods, and compliance criteria with international harmony in mind. This makes it easier for regulators and applicants in different countries to develop and approve products, respectively, and minimizes friction involved in recognizing authorizations across international boundaries.

Although airworthiness standards are commonly harmonized, acceptable methods of compliance may not always be, particularly where regulators have differing opinions over the criticality of a safety issue. As a result, situations sometimes arise where one regulator finds a design feature complaint with an airworthiness standard and another regulator may not because, although the standard is the same, the second regulator may require a different, more rigorous method of showing compliance that the applicant cannot satisfy.

V. Additional Considerations

Multiple methods of compliance. With respect to showings of compliance in design projects, sometimes tests may not be required, or tests may be used in conjunction with analysis. For example, regulations may require that components must be sufficiently dissimilar to prevent a single common-cause failure that propagates across redundant systems. An applicant might show compliance with such a requirement by: (i) demonstrating that the design prevents propagation of a common-cause failure; (ii) showing through quantitative or qualitative analysis that occurrence of such an event is sufficiently unlikely; or (iii) showing through testing of components that their failure rate will prevent occurrence. In a sophisticated certification project, it would not be unusual for all three methods to be used to demonstrate compliance.

Evolution of standards. Airworthiness standards are not static and are continually revised based on real world events. For example, for decades regulations specified minimum standards to guard against a catastrophic in-flight event. The explosion of TWA Flight 800, which was attributed to a chain of causes involving conductive fluid spilled on chafed wiring, the introduction of high electrical voltage into the fuel quantity measuring system, and a resulting spark in the main fuel tank, resulted in a comprehensive rewrite of regulations and guidance pertaining to, *inter alia,* electrical wiring systems, requirements for design features to prevent fuel system ignition, and the acceptable analysis for demonstrating that a design sufficiently mitigates risk of a catastrophic event. Pertinent standards have evolved further still, building on changes issued after the incident to encompass ignition-prevention more generally. For example, the FAA has looked at protection against adverse effects from lightning strikes and evolved its requirements to more specifically address mitigating risk of fuel tank ignition caused by a lightning strike.

This type of requirement is an example of an outcome-oriented standard augmented by industry consensus guidance. The pertinent regulation states that design and installation of a fuel system must prevent catastrophic fuel vapor ignition due to lightning and its effects. Industry, working through SAE International with the input of regulators, developed technical data and guidance describing methods of assessing lightning effects and mitigating ignition risks that the FAA accepts in connection with showings of compliance to the regulation.

Design changes. Because airworthiness standards evolve, questions sometimes arise over what standard applies, and what method of compliance will be acceptable, when an applicant modifies an existing approved design. For example, an aircraft manufacturer may want to significantly change an aircraft that was approved decades earlier in order to incorporate a new design feature. Regulations require the applicant to demonstrate that the new feature and aspects of the previously approved aircraft changed by it comply with the latest version of applicable airworthiness standards unless compliance with the latest version would not contribute materially to safety or would be impractical. This requirement sometimes results in disputes between applicants and regulators over whether the earlier or more recent airworthiness standard should apply.

Human factors considerations. Human factors criteria are usually not spelled out in specific airworthiness standards, but by policy FAA considers them in making findings of compliance. It relies on data and science developed by regulators, industry, and academia to consider whether a design feature is sufficiently safe and usable from a human factors standpoint. Depending on the project, the FAA may require an applicant to develop a human-factors specific certification plan to address such matters.

VI. Lessons Learned and Recommendations

Although industry has lobbied for and continues to support performance-based regulatory standards over prescriptive ones, in practice it has found that prescriptive standards have their place because they can foster certainty, repeatability, and international harmonization that substantially benefits industry. Even where standards are prescriptive, industry and regulators rely substantially on criteria, compliance processes, and testing methods developed by consensus through industry organizations, often with participation of regulators. This approach has proven successful. It supports innovation and evolution in means of compliance to address new technologies while affording a degree of repeatability and certainty in certification projects. Flexibility is further enhanced because individual applicants are generally not required to rely on industry standards if, on their own, they develop alternate means of showing compliance that are acceptable to regulators.

A significant downside to standards and means of compliance developed through industry consensus results from the potential need for applicants to share confidential and proprietary data with their competitors. As a result, development of consensus standards for novel and innovative technologies may be delayed or put on indefinite hold due to reluctance of some companies to provide their competitors with visibility into their trade secrets.

The mature, harmonized international aviation regulatory framework, or elements of it, could offer a helpful model for developers of standards and testing regimes in other domains. Fundamentally, testing criteria and methods are closely aligned with, and developed in concert with, the prevailing standards that apply to aviation products. The international regulatory regime is marked by a high degree of collaboration between and among countries, regulators, and applicants. It applies a safety continuum with respect to certification requirements, in which the rigor of requirements increases as risk increases, thereby establishing very strict standards for high-risk features, while offering increasing latitude respect to items that are not essential to safety. It embraces inputs from industry participants in the development of standards, testing methods, and means of showing compliance. It also supports the international recognition and interoperability of standards and testing processes across jurisdictions.

Moreover, activities in the aviation domain are largely defined by process. Industry participants embrace the need for consistent, repeatable, process-oriented approaches to their operations. This process orientation, which has been copied to varying degrees by other industries, forms the foundation for coordination and alignment among stakeholders, and is reflected in aviation's remarkable safety record. Indeed, it would be inaccurate to conclude that the comprehensiveness of the international regulatory regime stifles innovation in aviation products. To the contrary, the rigorous process orientation of the regime supports innovation. This is because activities of applicants are defined through detailed processes over which regulators have visibility and oversight. Adherence to process forms a foundation for safety and repeatability which supports innovation in the substance of the work. This is demonstrated by the nimble and rapid iteration currently taking place in novel technologies such as drone delivery systems and air taxis.

Aviation's comprehensive process orientation forms a basis for trust between regulators and applicants. If a regulator has confidence in an applicant's processes and systems for adhering to them, the regulator will have less incentive to resort to highly prescriptive regulations or vigorous, potentially suffocating, oversight. Learnings from aviation's comprehensive process orientation could be useful in development of testing and coordination activities as AI evolves on an international scale. For example, stakeholders could work toward alignment on processes for developing and iterating on testing methods, criteria, surveillance, and oversight before working through the substance of each such issue.