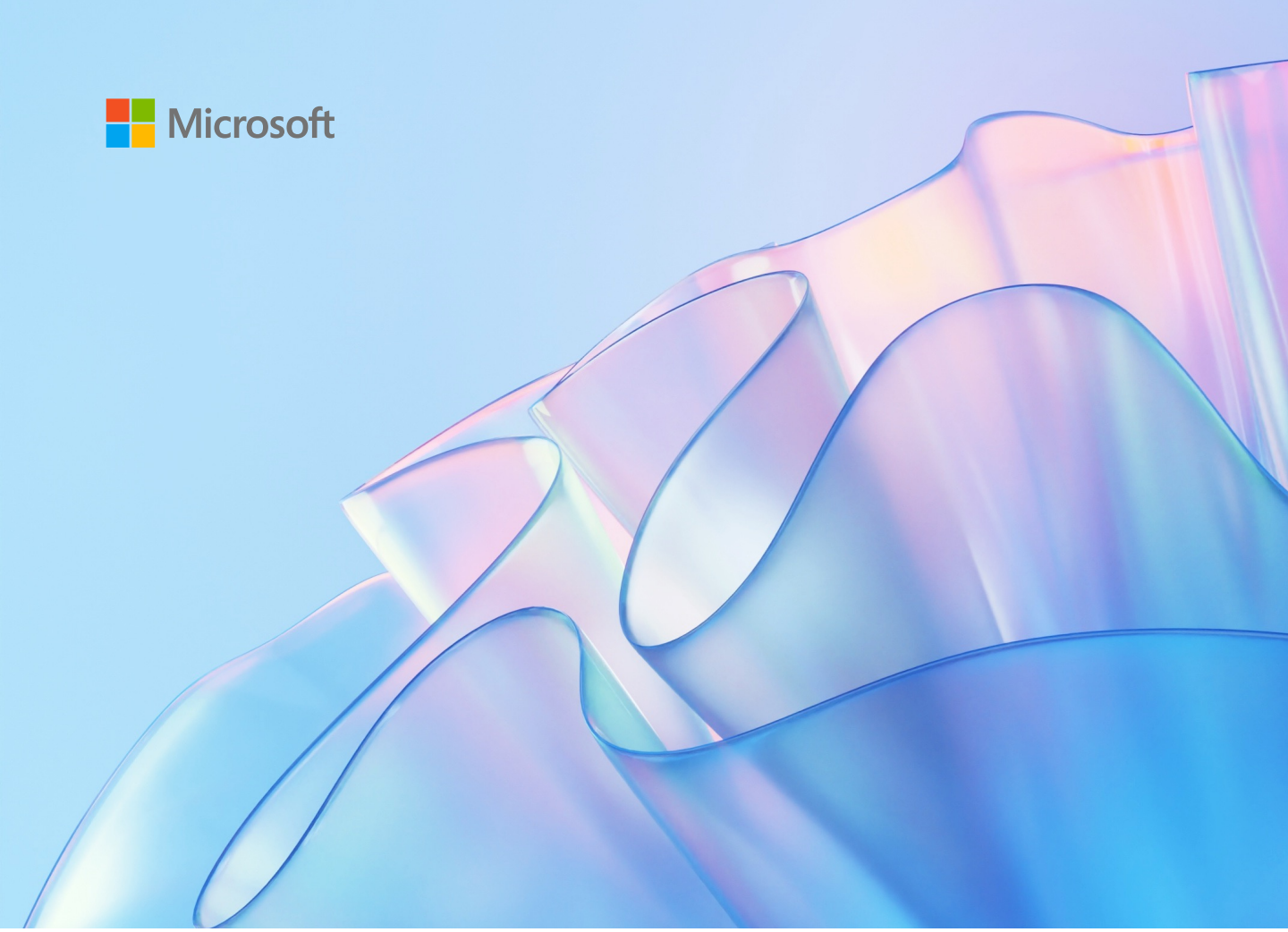
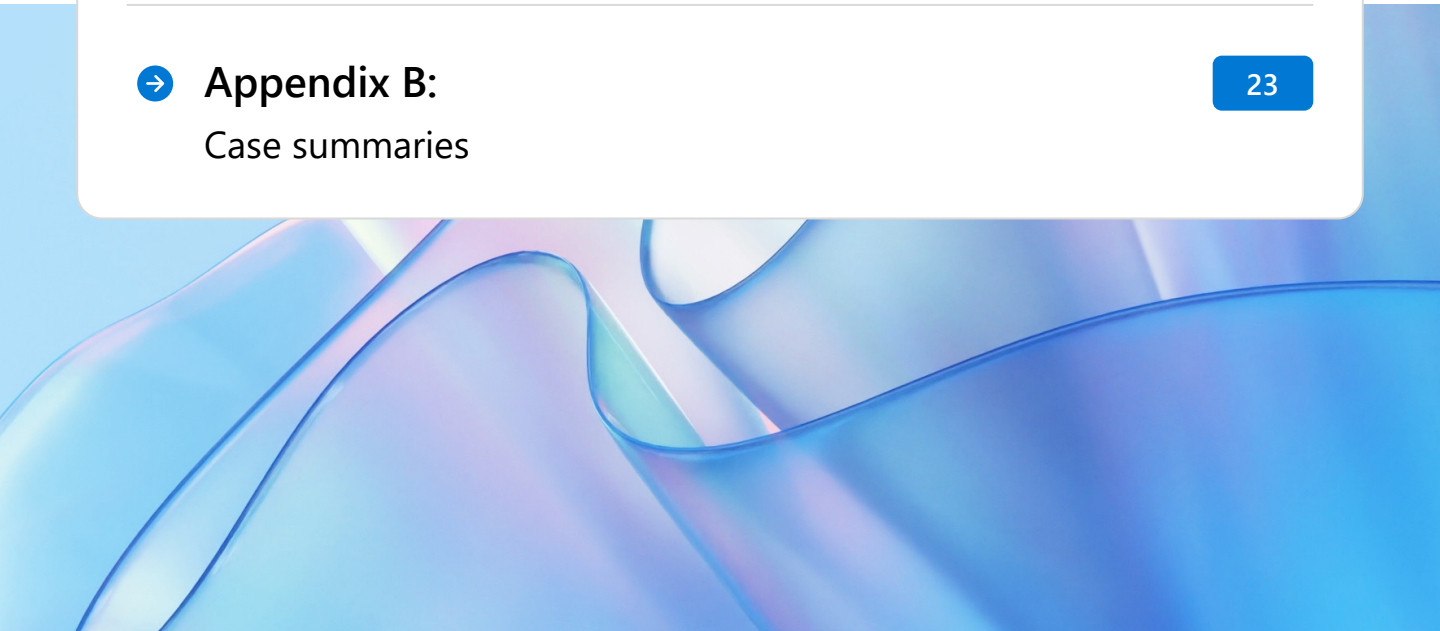


Learning from other domains to advance AI evaluation and testing

Table of contents



Introduction

As generative AI becomes more capable and widely deployed, familiar questions from the governance of other transformative technologies have surfaced. Which opportunities, capabilities, risks, and impacts should be evaluated so they can be measured and better understood? Who should conduct evaluations, and at what stages of the technology lifecycle? What tests and measurements should be used? How can we know if the results are reliable?

There is also growing awareness that evaluation of generative AI is more complex than evaluation of traditional machine learning systems.[1] This complexity is impacting the development of a rigorous science and practice of AI evaluation, which has been criticized in the New York Times as a 'tangle of sloppy tests [and] apple-to-oranges comparisons.' The International AI Safety Report (2025) (IASR)—the world's first comprehensive synthesis of research on the capabilities and risks of advanced AI, backed by 33 national governments— warns that the lack of rigorous and clear standards for risk evaluation creates 'an urgent policy challenge',[2] underscoring the need to strengthen AI evaluations as an 'integral' part of effective AI risk management.[3] America's AI Action Plan (July 2025) calls on United States (US) government agencies to advance 'the science of measuring and evaluating AI models' and recognizes the importance of building an AI evaluations ecosystem.[4]

This landscape motivated us to better understand how evaluation ecosystems have taken shape in other domains, with best practices, standards, and public policies co-evolving to support more ready and reliable ecosystems. Our research began in late 2024, when Microsoft's Office of Responsible AI gathered independent experts from civil aviation, cybersecurity, financial services (bank stress testing), genome editing, medical devices, nanotechnology, nuclear power, and pharmaceuticals. In bringing this group together, we built on insights and feedback from earlier cross-domain research, which culminated in the May 2024 e-book *Global Governance: Goals and Lessons for AI*. That publication featured expert case studies on international institutions addressing cross-border issues—including in civil aviation, nuclear energy, and global finance—and drew lessons from their high-level goals and governance approaches. Following that effort, we decided to go deeper on a specific aspect of governance—evaluation and testing—given its growing salience in AI policy. We launched eight case studies, bringing in domains suggested during discussions about our e-book and expanding our focus on general-purpose technologies (nanotechnology and genome editing).

We identified several key variables that shape differences in how each of the eight domains developed evaluation ecosystems and the role evaluation and testing play in governance. These include the types of technologies, deployment contexts, and risk profiles at issue; the historical moments in which the evaluation, testing, and assurance frameworks were developed and later solidified; the maturity of the evaluation science and the nature

[1] Wallach, H., et al. (2025). *Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge.* https://arxiv.org/abs/2502.00561.
[2] Section 3.3, page 181.
[3] Page 23.
[4] Pillar I, page 10.

of the stakeholder communities involved in advancing it; and the placement of expertise in the assessor ecosystem. These variations and the independent experts' analyses of their impacts informed six key takeaways relevant to AI evaluation, testing, and governance:

- **First, testing is a cornerstone of trust in critical systems,** enabling stakeholders across domains to evaluate whether technologies, medicines, aircraft, or even financial institutions will perform as expected and avoid consequential unintended side effects.
- **Second, effectively embedding testing within governance frameworks requires addressing foundational questions** about what is tested, how tests are conducted, and how results are used. Mature frameworks rely on rigor in defining what is being tested and why; standardization of how tests should be conducted to achieve reliable results; and a clear understanding of how to interpret test results and use them to inform decisions about technology deployment and risk management investments. An understanding of the benefits and limitations of testing underpins these principles. Where these scientific, methodological, and procedural foundations are strong, testing can be leveraged more effectively as a governance tool.
- **Third, public policy frameworks for evaluation and testing reflect trade-offs among governance objectives, such as safety, efficiency, and innovation.** Experts from all eight domains noted that policymakers have had to weigh trade-offs in designing policy frameworks. Moreover, frameworks have had to account for both the limits of current science and the need for agility in the face of uncertainty. Experts likewise agreed that early design choices, often reflecting the "DNA" of the historical moment in which they're made,[5] are important as these decisions have proven difficult to scale back or reverse later.
- **Fourth, trade-offs hold most firmly in strict, pre-deployment testing regimes,** such as those used in civil aviation, medical devices, nuclear power, and pharmaceuticals. Testing regimes in these domains offer strong pre-deployment safety assurances but can be resource-intensive and slow to adapt, and at least in one domain,[6] they also result in less emphasis on post-deployment monitoring, which measures real-world impact of safety and security measures. Strict pre-deployment testing regimes have often emerged in response to well-documented failures and are backed by decades of regulatory infrastructure and detailed technical standards.
- **Fifth, more adaptive governance frameworks tend to be utilized in domains marked by more rapid technological change and dynamic interactions with the external environment**—such as cybersecurity and bank stress testing. In these domains, testing is used to generate actionable insights about risk, with relatively less emphasis on the role of testing in pre-deployment regulatory authorization. Pharmaceuticals provides a counterexample, as a domain with complex and potentially dynamic interactions between a tested system and its deployment environment (e.g., people with variable drug reactions).
- **Sixth, the most adaptive frameworks apply to general-purpose technologies (GPTs), where the contexts in which they're deployed vary most widely.** In these domains,

[5] Baker, S. (2025). *Cybersecurity Case Study.*
[6] Carpenter, D. and Benamouzig, D. (2025). *Pharmaceuticals Case Study.*

including genome editing and nanotechnology, evaluation and testing can apply "upstream" to the GPT or "downstream" in specific deployment contexts. While upstream evaluation and governance may be perceived as offering efficiencies or stronger risk avoidance, too much emphasis upstream can also limit responsiveness to the varied opportunities, capabilities, risks, and impacts that emerge downstream. In genome editing, for example, jurisdictions that regulate based on downstream effects are often more permissive of genome-edited crops where no foreign DNA is introduced.[7] In contrast, jurisdictions that target governance upstream tend to apply stricter rules to such genome-edited crops—even if experts suggest their risk level does not require it—treating them similarly to transgenic genetically modified organisms (GMOs) that incorporate foreign DNA from other species.[8]

These takeaways can inform next steps to advance an AI evaluations ecosystem, which will require addressing gaps in science and practice. While various methods for evaluating AI models and systems have emerged (several of which are defined in Appendix A), experts leading efforts to design and implement AI evaluations at the technical frontier have identified challenges with existing methods.

- The IASR finds that current quantitative methods for assessing general-purpose AI have 'significant limitations' and that evaluators working at the technical frontier 'need substantial and growing technical ability and expertise.'[9] It especially calls attention to challenges with the 'validity' and 'reliability' of measurement techniques (i.e., accuracy and 'consistency, stability, and dependability of a measurement over time and across different contexts').[10] AI model benchmark evaluations (defined in Appendix A), for example, are widely utilized but may lose effectiveness over time. As new models rapidly improve, benchmarks may become saturated, with most models scoring at the test's maximum, making it harder to detect and compare marginal progress in outcomes of interest across time. Their reliability is further undermined when openly published benchmarks are incorporated into AI model training data, resulting in models essentially having "answers" to the test.

- In practice, generative AI systems[11] produce outputs that can vary from one session to another, even when given the same input. This variability arises from several sources, including probabilistic sampling during generation, sensitivity to prompt phrasing and system instructions, and context from prior messages. As a result, model outputs exhibit stochastic behavior: patterns can be analyzed statistically across multiple runs, but any single output is difficult to predict with precision. This poses challenges for performance benchmarking, as evaluations conducted in one context may fail to reproduce results in another, unless all variables—including system prompts, model versions, and runtime parameters—are tightly controlled.[12]

- The IASR further observes that existing evaluations 'mainly rely on 'spot checks', i.e., testing the behaviour of a general-purpose AI in a set of specific situations.'[13]

[7] Charo, A. and Greenfield, A. (2025). *Genome Editing Case Study.*
[8] Ibid.
[9] Section 3.3, page 181.
[10] Section 3.3, page 184.
[11] In this white paper, we adopt the IASR definition of "system": '[a]n integrated setup that combines one or more AI models with other components, such as user interfaces or content filters, to produce an application that users can interact with.' Page 201.
[12] Blackwell, R. E., et al. (2024). *Towards reproducible LLM evaluation: Quantifying uncertainty in LLM benchmark scores.* https://arxiv.org/pdf/2410.03492.
[13] IASR, Executive Summary, page 23.

When evaluating general-purpose models, it is challenging for evaluators to cover the large spectrum of AI capabilities and potential impacts. While spot checks can help surface potential hazards, they may 'miss hazards and overestimate or underestimate' capabilities and risks,[14] undermining their ability to produce actionable insights for countless potential downstream AI applications.

- Building on the IASR, the Singapore Consensus (2025)—which captures research priorities shared by more than 100 global experts—emphasizes the need to prioritize research into AI risk assessment and to develop standardized approaches for measuring the impacts and behaviors of current and future AI systems.[15]
- More broadly, research has highlighted how designing measurement instruments, such as benchmark tests, without systematically examining the complex, nuanced, and sometimes contested concepts that they aim to measure, can result in instruments that do not accurately reflects complexities;[16] how evaluating language-based models in one language does not always approximate capabilities and risks in another (this is especially so when languages utilize different scripts, such as Latin (English) or Devanāgarī (Hindi));[17] and that more education is needed to help stakeholders interpret AI evaluation results, often presented as performance leaderboards or narrowly framed metrics that can obscure nuanced, qualitative insights into a general-purpose model's suitability and reliability for different applications.[18]
- Moreover, rapid advances in AI, such as the emergence of agentic AI systems, offer exciting new capabilities but also demand novel approaches to evaluation.[19] Like traditional software, agentic AI systems require evaluations for task completion, efficiency, reliability, and unintended side-effects. However, as they also may take action in a more dynamic and interactive environment, evaluations must be tailored to agentic AI systems' specific tasks and simulated environments, including potential adversarial conditions such as deceptive pop-ups aiming to elicit sensitive information.

Drawing on our analysis of eight case studies prepared by independent academic and industry experts, this white paper[20] proposes next steps to address AI evaluation and testing challenges and opportunities by:

- Synthesizing **insights from the eight case studies,** also published separately, and extracting lessons relevant to AI (Part 1);
- Surveying **key multistakeholder initiatives** that are driving AI evaluation science and practice forward (Part 2); and
- Presenting **recommendations for policymakers** aiming to advance the AI evaluation and testing ecosystem and strengthen AI governance (Part 3).

[14] Ibid.
[15] Pages 9 – 14.
[16] IASR, Section 3.3; Wallach, H., et al (2025). *Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge.* https://arxiv.org/pdf/2502.00561.
[17] Presentation by Sunayana Sitaram, Principal Researcher at Microsoft Research India, to AI Safety and Security Institute Network (April 2025).
[18] Burden, J. (2024). *Evaluating AI Evaluation: Perils and Prospects.* https://arxiv.org/pdf/2407.09221.
[19] AI Security Institute. (July 2025). *International joint testing exercise: Agentic testing* (blog). https://www.aisi.gov.uk/work/international-joint-testing-exercise-agentic-testing.
[20] We thank the Oxford Martin AI Governance Initiative for hosting a roundtable at the University of Oxford, where we presented early findings from this research and received helpful feedback. We also thank MLCommons and Patricia Paskov for their helpful feedback on earlier drafts of this white paper (all errors are our own). In addition, we are grateful to the following external experts who contributed to our cross-domain research program on lessons for AI evaluation and testing: Mateo Aboy, Paul Alp, Gerónimo Poletto Antonacci, Stewart Baker, Daniel Benamouzig, Pablo Cantero, Daniel Carpenter, Alta Charo, Jennifer Dionne, Andy Greenfield, Kathryn Judge, Ciaran Martin, and Timo Minssen.

Part 1: Lessons learned from studying eight domains

Evaluation and testing play a central role or act as significant governance tools across the eight domains examined by Microsoft. To understand the role of evaluation and testing across these domains, we asked several foundational questions that are also being explored in the context of AI governance. Which opportunities, capabilities, risks, and impacts should be evaluated? How should risk thresholds be set? Who should conduct evaluations, and at what stages of the technology lifecycle? What tests and measurement instruments should be used? How can we know if the results are reliable? And what role should evaluation and testing play within governance and public policy frameworks? Appendix B summarizes how each of the eight domains approach these questions.

In pharmaceuticals, medical devices, civil aviation, and nuclear technology, testing is the backbone of a strict regime of pre-deployment regulatory approval. In pharmaceuticals, regulators such as the Food and Drug Administration (FDA) and European Medicines Agency (EMA) assess whether a drug is safe for market by considering the results of rigorous, multi-phase controlled trials.[21] In civil aviation, 'detailed requirements' are comprehensively imposed on the design, manufacture and operation of aircrafts.[22] Multiple types of testing, underpinned by detailed technical standards, are undertaken to demonstrate the airworthiness of designs, aircrafts, and their componentry.[23] These strict regimes emerged in connection with historically significant incidents that demonstrated the dangers of faulty drugs, such as the thalidomide tragedy (1960s); unsafe aircraft designs, such as the crash of the Comet (1954); and nuclear accidents, such as Chernobyl (1986) and Fukushima (2011).

Despite well-documented failures, there are ongoing debates about whether strict pre-deployment testing regimes strike the right balance between safety, efficiency, and innovation. In pharmaceuticals, only the 'largest global biopharmaceutical firms can invest in large-scale evaluation processes…'[24] These cost 'hundreds of millions or even billions of dollars' and run the risk that the FDA or EMA will reject new drug applications.[25] These dynamics may undermine market competition and innovation. Researchers, pharmaceutical companies, patients, and advocacy groups have called for relaxing clinical trial requirements—particularly for pre-deployment testing—recognizing that an excessive emphasis on safety over efficiency can limit the availability of potentially life-saving treatments.[26] However, unravelling such a deeply entrenched regime is challenging, as early design choices—often reflecting the historical context in which they were made—are hard to unwind later.

Strict regimes—such as pharmaceuticals, medical devices, civil aviation, and nuclear power—rely on strong scientific foundations and detailed technical standards for testing. These foundations evolved over time due to multiple inputs, including scientific

[21] Carpenter, D. and Benamouzig, D. (2025). *Pharmaceuticals Case Study.*
[22] Alp, P. (2025). *Civil Aviation Case Study.*
[23] Ibid.
[24] Carpenter, D. and Benamouzig, D. (2025). *Pharmaceuticals Case Study.*
[25] Ibid.
[26] Ibid.

innovation, multi-year standardization efforts, and international coordination and information sharing around best practices and lessons learned from incidents. In pharmaceuticals, an approval regime 'evolved due to scientific innovation, public and economic demand for information about products, and political pressure forged through critical historical episodes.'[27]

In nuclear power, decades of experience have shaped rigorous safety protocols, with testing evolving from basic standards into a comprehensive, performance-based governance system that incorporates insights from historical incident analysis, operational experience, and advances in safety science.[28] The International Atomic Energy Agency (IAEA), which Microsoft examined as part of *Global Governance: Goals and Lessons for AI*, has been central to building shared understanding among the international community about 'overarching core testing requirements' for nuclear power.[29] This has been a multidecade process. Even with these sustained efforts, new innovations, such as small modular reactors, mean that comprehensive harmonization of standards is 'a distant goal.'[30]

Dynamic, exploratory approaches to testing and more adaptive governance frameworks emerge in domains that are contending with rapid change and interactions with complex deployment environments. In such domains, the relevance of technical standards and best practices may quickly shift. This is evident in cybersecurity, where complexity and the rapid pace of technical change make setting 'hard and fast' rules challenging and impact enforcement.[31] The strict regimes discussed above set lengthy and prescriptive regulatory guidance, including about which risks and impacts should be evaluated, and, in some instances, what tests or measurements should be used. By contrast, cybersecurity policymakers have, broadly speaking, set flexible standards defined by reference to continuously evolving industry best practices.[32] Policymakers may, however, mandate technical controls, such as encryption or multi-factor authentication,[33] while allowing for flexibility in implementation approaches.

Where risks have multiple causes and result from complex interactions in the real-world, the predictive power of ex ante (or pre-deployment) evaluations is inherently limited. This is evident in the banking sector, where a process of 'stress testing' banks has emerged since 2008.[34] Stress testing of banks 'entails assessing how that bank will fare under a given adverse scenario, usually meant to replicate the types of developments that would occur during a deep or prolonged recession...'[35] Despite meaningful improvements over time, stress tests struggle to capture dynamic feedback loops that arise during financial crises—such as liquidity hoarding, contagion effects, fire sales, fear, and dysfunction in secondary markets—due to the complexity and unpredictability of these variables.[36] As a result, even well-designed banking stress tests may underestimate systemic vulnerabilities. In recognition of these limitations, regulators like the US Federal Reserve have introduced exploratory analyses, which test 'how banks and the banking system would fare under a

[27] Ibid.
[28] Cantero, P. and Poletto Antonacci, G. (2025). *Nuclear Power Case Study.*
[29] Ibid.
[30] Ibid.
[31] Baker, S. (2025). *Cybersecurity Case Study.*
[32] Baker, S. (2025). *Cybersecurity Case Study;* Microsoft. (2025). *AI Testing and Evaluation: Learnings from cybersecurity* (Podcast). Episode 3.
[33] Ibid.
[34] Judge, K. (2025). *Bank Stress Testing Case Study.*
[35] Ibid.
[36] Ibid.

greater range of scenarios for purely informational purposes.'[37] Decoupling these exploratory analyses from specific capital adequacy requirements makes it easier for the Federal Reserve to be creative in its analysis and uncover new insights into the resilience of the banking system and sources of fragility.[38]

Governance of general-purpose technologies (GPTs), such as genome editing tools and nanotechnology, requires strategic calibration of testing and oversight upstream—on the GPT—or downstream, at the level of specific applications. Genome editing[39] and nanotechnology[40] illustrate how risks tend to become more visible and assessable once the technology is applied to a particular use case and context-specific variables, such as exposure pathways and affected populations, can be clearly defined.

Whereas the EU has regulated genome editing more horizontally upstream, the US has largely adopted an application-specific approach to genome editing governance, applying different regulatory standards depending on whether the technology is used in agriculture, medicine, or other sectors.[41] Professor Alta Charo (University of Wisconsin-Madison) explained that the US approach is based on novelty and risk: truly new or potentially risky use cases (i.e., drugs) require prior approval and pre-market controls, while familiar or low-risk use cases can proceed directly to market with failures remediated through post-market controls.[42] In the EU, horizontal regulation has meant that deployment and innovation may have lagged for use cases that other jurisdictions have interpreted as lower risk.[43] Applying uniform rules to genome editing results in a framework that is precautionary but rigid, potentially stifling innovation unnecessarily for well-understood scenarios and low-risk use cases.[44]

In nanotechnology, there is 'relatively little [horizontal] regulation governing standardization of nanomaterials use and safety testing.'[45] In certain sectors where the application of nanomaterials is low-risk, '[l]ighter-touch governance… is both deliberate and desirable.'[46] For example, the use of nanomaterials as additives in air and water filtration systems or industrial manufacturing is not subject to strict regulation, as these materials are considered to be adequately 'packaged', preventing direct interaction by the end-user.[47] By contrast, in the pharmaceuticals sector—where lipid nanoparticles are used for drug delivery across a range of conditions, including autoimmune disorders, cancer, metabolic diseases, and infectious diseases, as well as in prophylactics and vaccines—nanoparticles are subject to rigorous testing through multiple phases of clinical trials before being approved for use in humans.[48] While a vertical approach to governance can introduce uncertainty in novel cases, it brings efficiencies in well-understood cases and allows risk assessment and management to be more responsive to real-world deployment contexts.

[37] Ibid.
[38] Ibid.
[39] Charo, A. and Greenfield, A. (2025). *Genome Editing Case Study.*
[40] Dionne, J. (2025). *Nanoscience and Nanotechnology Case Study.*
[41] Charo, A. and Greenfield, A. (2025). *Genome Editing Case Study.*
[42] Microsoft Research. (2025). *AI Testing and Evaluation: Learnings from genome editing* (Podcast). Episode 1.
[43] Charo, A. and Greenfield, A. (2025). *Genome Editing Case Study.*
[44] Ibid.
[45] Dionne, J. (2025). *Nanoscience and Nanotechnology Case Study.*
[46] Ibid.
[47] Ibid.
[48] Ibid.

GPTs reveal a tension between developing testing methods and tools that are broadly generalizable and ensuring they remain responsive to the specific characteristics of downstream applications. This tension also complicates the standardization of norms for how evaluation and testing should be conducted.

In nanotechnology, for instance, the specialized characterization equipment required for thorough evaluation of nanomaterials is difficult to scale.[49] As Professor Jennifer Dionne (Stanford University) notes, 'such tools are so specialized, linking the (often heterogeneous) nanoscale composition of nanomaterials with downstream function in applications spanning batteries, catalysis, optoelectronic devices, and even biomedical devices is largely underexplored.'[50] As a result, the semiconductors and pharmaceuticals sectors have developed sector-specific inspection tools such as those used to inspect wafers and chips or to ensure quality control of drug manufacturing. However, scaling such tools for routine risk assessments across the diverse sectors using nanomaterials would be complex, cost-prohibitive and 'reveal a scarcity of expertise in nanocharacterization.'[51]

Post-deployment monitoring is an example across domains of assessing how a product performs downstream in the real-world context of its deployment. Post-deployment monitoring takes different forms across domains and is formalized in governance in a variety of ways. In medical devices, both the EU and US mandate adverse event reporting and regular surveillance to monitor device performance post-deployment.[52] The expansion of testing requirements to ongoing monitoring and evaluation emerged in the 1990s, '[underscoring] the idea that a device's safety and efficacy could evolve post-approval, necessitating continual testing and data collection…'[53]

In cybersecurity, practices like coordinated vulnerability disclosure, security researcher recognition programs, and bug bounties have emerged as key to ongoing efforts to strengthen security posture. While their value has been reflected in cybersecurity risk management frameworks and standards,[54] the practices emerged and solidified over more than a decade of norm building between technology companies, security researchers, and operational coordinators[55] in advance of being incorporated in regulation.[56] This process has meant improvements from this form of ongoing monitoring have been driven by operational practice. One illustrative example is bounty programs, which reward researchers for finding flaws in software that has already been deployed. Microsoft's CodeQL system takes findings a step further: a confirmed vulnerability is translated into a query that scans for similar issues across Microsoft's entire codebase. CodeQL treats code as data, enabling automated variant analysis to find and fix systemic issues rather than individual bugs. While AI lacks a neat issue–patch model, this culture of iterative discovery, risk mitigation, and scaled mitigation via tooling offers a useful analogy. Post-deployment monitoring

[49] Ibid.
[50] Ibid.
[51] Ibid.
[52] Aboy, M. and Minssen, T. (2025). *Medical Devices Case Study.*
[53] Ibid.
[54] National Institute of Standards and Technology. (2024). *The NIST Cybersecurity Framework (CSF) 2.0.* (National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 29. https://doi.org/10.6028/NIST.CSWP.29; International Organization for Standardization and International Electrotechnical Commission. (2018). *Information technology — Security techniques — Vulnerability disclosure* (ISO/IEC Standard No. 29147:2018); International Organization for Standardization and International Electrotechnical Commission. (2019). *Information technology — Security techniques — Vulnerability handling processes* (ISO/IEC Standard No. 30111:2019).
[55] E.g., Microsoft Security Response Centre. (2010). *Coordinated Vulnerability Disclosure: Bringing Balance to the Force* (blog). https://msrc.microsoft.com/blog/2010/07/coordinated-vulnerability-disclosure-bringing-balance-to-the-force/; NTIA Awareness and Adoption Group. (2016). *Vulnerability Disclosure Attitudes and Actions.* https://www.ntia.gov/files/ntia/publications/2016_ntia_a_a_vulnerability_disclosure_insights_report.pdf.
[56] The European Cyber Resilience Act.

of AI[57] may similarly evolve into a process of surfacing issues, generalizing them, and systematically mitigating risks over time.

Where there is a regulatory emphasis on pre-deployment testing, experts highlighted a potential trade-off with post-deployment monitoring. This is particularly the case for pharmaceuticals. Though different types of testing occur across the lifecycle of a drug, testing is 'front-loaded', meaning that 'most of it occurs before regulatory authorization'.[58] Post-market studies of safety and efficacy—so-called "Phase IV commitments"—are carried out slowly or not realized at all.[59] This may be because '[p]harmaceutical sponsors have the greatest incentive to conduct costly experiments before the drug is authorized because it is the regulator, and not the company, who makes the final launch decision.'[60] This relates to an emerging tension in AI governance: while pre-deployment evaluations help assess whether a model or system is ready and reliable for release, post-deployment monitoring is essential to detect real-world risks that only surface under actual conditions of use. Effective policy and regulatory frameworks must aim to incentivize the right balance of both, while strengthening feedback loops so that post-deployment insights inform future pre-deployment testing and risk mitigation strategies.

Finally, the expert case studies suggest that while robust evaluation is essential to effective risk management, it is not sufficient on its own. Even with strong and appropriately calibrated pre-deployment evaluation and post-deployment monitoring, complex real-world dynamics can still give rise to unforeseen risks. In genome editing, for example, several complications confound risk assessments, including the 'absence of the comprehensive underlying data needed to predict specific effects', the 'inherently subjective nature' of evaluating certain effects, and the 'uneven distribution of effects within the population'.[61] These challenges are exacerbated if the edit is made in 'poorly understood areas of the genome.'[62] Likewise, in nanotechnology, there is 'potential for unintended nanomaterial formation', when, for example, bulk materials degrade into nanoparticles with highly variable properties.[63]

Similarly, in AI governance, even with an improved evaluation and testing ecosystem, there will be blind spots and limitations. This highlights the importance of a broader risk management approach that is adaptive, informed by transparency, and supported by feedback loops and continuous learning throughout the AI development and deployment lifecycle. In civil aviation, which enjoys a mature and internationally harmonized governance ecosystem, these objectives have been facilitated by 'a high degree of collaboration between and among countries, regulators, and [industry]'.[64]

In the next section, we consider the emerging role of AI evaluation in public policy and survey recent multistakeholder initiatives in advancing the AI evaluation and testing ecosystem.

[57] Cattell, S., et al. (2024). *Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities.* In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Vol. 7, pp. 267-280); Stein, M. and Dunlop, C. (2024). Safe beyond sale: post-deployment monitoring of AI (blog). *Ada Lovelace Institute.* https://www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/.
[58] Carpenter, D. and Benamouzig, D. (2025). *Pharmaceuticals Case Study.*
[59] Ibid.
[60] Ibid.
[61] Charo, A. and Greenfield, A. (2025). *Genome Editing Case Study.*
[62] Ibid.
[63] Dionne, J. (2025). *Nanoscience and Nanotechnology Case Study.*
[64] Alp, P. (2025). *Civil Aviation Case Study.*

Part 2: Ongoing initiatives to advance the AI evaluation and testing ecosystem

Policymakers worldwide are embedding AI risk assessment and evaluation as an expectation or requirement in emerging policy and regulation.[65] New York City (NYC) established an algorithmic auditing regime as early as July 2023, focusing on the use by NYC-based employers of automated employment decision-making tools.[66] Scholars found that one of several gaps hampering the implementation of the NYC law was the absence of technical standards, including 'methodological details surrounding the approach to [AI] auditing...'[67] Numerous other examples have emerged globally:

- The European Union's AI Act requires testing of both general-purpose AI (GPAI) models and high-risk AI systems. Providers of 'general-purpose AI models with systemic risk' are required to 'assess and mitigate possible systemic risks,' performing model evaluations 'through internal or independent external testing,' with the GPAI Code of Practice Safety and Security Chapter defining more detailed expectations.[68] The Act also requires 'high-risk AI systems' to undergo testing as part of quality management[69] and for evaluation of risk throughout an AI system's lifecycle.[70]
- South Korean AI legislation requires 'AI Business Operators' to conduct risk assessments if the models powering their AI systems were trained using computational resources beyond a certain threshold.
- In the United States, federal agencies are required to conduct pre-deployment testing for high-impact AI.[71] Among US states, including California[72] and New York,[73] bills are being considered that would require developers of frontier AI models to describe their testing procedures and safety protocols as well as report the high-level results of risk assessments. The proposed California bill would also require developers to describe the extent to which they use assessments performed by independent third parties, without mandating such assessments.

Voluntary AI standards also emphasize the importance of evaluation and testing. For example, under its 'measure' function, the NIST AI Risk Management Framework (NIST AI 100-1) states that 'AI systems should be tested before their deployment and regularly while in operation.'[74] Similarly, in Australia, the federal government's Voluntary AI Safety Standard recommends that deployers '[t]horoughly test AI systems and AI models... and then monitor for potential behaviour changes or unintended consequences.'[75]

[65] Burden, J., et al. (2025). *Paradigms of AI Evaluation: Mapping Goals, Methodologies and Culture.* https://arxiv.org/abs/2502.15620.
[66] NYC Local Law 144.
[67] Groves, L., et al. (2024). Auditing work: Exploring the New York city algorithmic bias audit regime. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1107-1120).
[68] Article 55, Recital 114, Article 92.
[69] Article 17.
[70] Article 9.
[71] Vought, R. T. (April 2025). *Memorandum for the Heads of Executive Departments and Agencies: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust* (Memorandum No. M-25-21). Executive Office of the President. Office of Management and Budget. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf.
[72] California Legislative Information. *SB-53 Artificial intelligence models: large developers* (bill text). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53.
[73] The New York State Senate. *Assembly Bill A6453A* (bill text). https://www.nysenate.gov/legislation/bills/2025/A6453/amendment/A.
[74] NIST Trustworthy and Responsible AI - NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1.
[75] Guardrail 4.

The prominence of AI evaluation and testing in governance frameworks reflects the foundational role of risk assessment in risk management frameworks across domains. However, the extent to which implementation is feasible and useful for actors across the AI supply chain—including AI model developers as well as AI system developers, deployers, and users, ranging from citizens to government agencies and companies across sectors—will be central to the effectiveness of these measures.

Private sector and multistakeholder initiatives have emerged to adopt and practically advance AI evaluation and testing, including through research and the development of best practices and tools. Collectively, these efforts represent early but meaningful steps toward a more reliable AI evaluation ecosystem.

- Leading model developers have committed to frontier AI safety policies, which lay out how they will assess their most advanced models for risks related to national security and public safety. Microsoft, for example, has published the Frontier Governance Framework.
- The Frontier Model Forum has published issue briefs and technical reports on current topics in frontier model safety evaluations. For example, a taxonomy (December 2024) helps distinguish methodologies and objectives of safety evaluations, and an early best practices brief (July 2024) offers recommendations related to the design, implementation, and disclosure of frontier AI safety evaluations. A Technical Report on Frontier Capability Assessments (April 2025) discusses emerging industry best practices for evaluating risks to public safety and security, such as advanced cyber threats.
- MLCommons, an open engineering consortium, is designing state-of-the-art and independent benchmarks. In December 2024, MLCommons launched AILuminate—a first-of-its kind, risk-focused benchmark for LLMs. Developed with contributions from leading research institutions and companies, AILuminate's benchmark assesses LLM responses to over 24,000 test prompts across twelve categories of physical, non-physical, and contextual hazards. In June 2025, MLCommons announced an effort to build a new agentic reliability evaluation standard, including frameworks and benchmarks addressing practical needs.[76]
- Technical committee SC 42 of the International Standards Organization (ISO) is developing a technical specification (ISO/IEC DTS 42119-2) that will describe testing techniques applicable to AI systems.[77]
- Microsoft researchers are advancing scholarship on AI evaluation frameworks. This includes the introduction of the ADeLe framework, which pioneers an evaluation technique that assesses how demanding a task is for an AI model by applying measurement scales for 18 types of cognitive and knowledge-based abilities. In addition, Microsoft researchers are developing new routes to greater systematization and more reliable AI tests. Examples include Eureka: Evaluating and Understanding Large Foundation Models (2024)[78] and Evaluating Generative AI Systems is a Social Science

[76] MLCommons. (June 2025). *MLCommons Builds New Agentic Reliability Evaluation Standard in Collaboration with Industry Leaders* (blog). https://mlcommons.org/2025/06/ares-announce/.
[77] International Organization for Standardization and International Electrotechnical Commission. *Artificial intelligence — Testing of AI* (ISO/IEC Standard No. DTS 42119-2). https://www.iso.org/standard/84127.html.
[78] Balachandran, V., et al. (2024). *EUREKA: Evaluating and Understanding Large Foundation Models*. https://arxiv.org/pdf/2409.10566.

Measurement Challenge (2024).[79]

- Microsoft is also developing practical tools and methods to strengthen testing in real-world settings. These include the Python Risk Identification Toolkit for Generative AI (PyRIT), based on the work of our AI Red Team, and evaluation capabilities embedded in Azure AI Foundry, based on tools used internally and grounded in internal research.

Public sector-led initiatives are also driving progress on AI evaluation science as well as cross-border coordination and coherence.

- America's AI Action Plan (July 2025) (AI Action Plan) emphasizes the importance of building a robust AI evaluation ecosystem to support trustworthy and secure AI development. It recognizes that rigorous evaluations are essential for measuring and improving the reliability and performance of AI systems, particularly in regulated and high-stakes domains, and that evaluations may increasingly support the application of existing laws to AI. To advance this vision, the AI Action Plan recommends several policy actions,[80] including: publishing guidelines and resources through the National Institute of Standards and Technology (NIST), including via the Center for AI Standards and Innovation (CAISI), to support federal agencies in conducting mission-specific AI evaluations; supporting the development of the science of AI measurement and evaluation through efforts led by NIST, the Department of Energy (DOE), the National Science Foundation (NSF), and other federal science agencies; investing in AI testbeds to pilot AI systems in secure, real-world settings across sectors such as agriculture, healthcare, and transportation; and convening biannual meetings to enable knowledge-sharing between agencies and researchers.
- In parallel, the AI Action Plan underscores the need for the US federal government to lead in evaluating national security risks posed by frontier AI models, including cyber threats, chemical, biological, radiological, nuclear, and explosives (CBRNE) risks, and adversarial AI use in critical infrastructure. Recommended policy actions[81] include evaluating frontier models for security risks in partnership with developers and relevant federal agencies; assessing the risks associated with foreign AI systems deployed in the US economy; and building and maintaining national security-focused AI evaluations in collaboration with national security agencies and research institutions.
- A major focus of the network of AI Safety and Security Institutes and the US CAISI (Network), comprising leading technical experts from governments across the world, has been to build a shared understanding of general-purpose AI model evaluations and to define core principles for assessing the risks of advanced AI systems. As part of these efforts, the Network has conducted a series of joint testing exercises: UK and US experts conducted joint pre-deployment testing of OpenAI's o1 model, publishing findings in December 2024; experts from Singapore, Japan, Australia, Canada, the European Union, France, Kenya, South Korea and the UK conducted joint multilingual testing of the Mistral Large and Gemma 2 (27B) models across ten languages, publishing findings in June 2025; and experts from the same Network members conducted a third exercise to develop and experiment with testing methodologies for common and cybersecurity risks

[79] Munoz, G. D., et al. (2024). *PyRIT: A Framework for Security Risk Identification and Red Teaming in Generative AI Systems*. https://arxiv.org/pdf/2410.02828.
[80] America's AI Action Plan (July 2025), Pillar I, page 10.
[81] America's AI Action Plan (July 2025), Pillar III, page 22.

of agentic AI systems, publishing findings in July 2025.

Individual public institutes, centers, and other authorities are also publishing reference points that can serve a broader community.

- For example, the UK AI Security Institute maintains a publicly available repository of community-contributed benchmark evaluations for LLMs, and the Japanese AI Safety Institute published guidance on red-teaming methodologies in September 2024. The UK's Financial Conduct Authority (FCA) plans to launch 'AI Live Testing', which would see the FCA's AI Lab work directly with firms to provide support on methods for evaluating the impact of AI models used in financial markets.
- Singapore's AI Verify Foundation launched a Global AI Assurance Pilot in February 2025, pairing firms deploying generative AI applications with firms specializing in AI assurance testing to explore approaches and challenges and surface best practices.
- The AI Action Plan directs the US Department of Commerce to convene the NIST AI Consortium to 'empower the collaborative establishment of new measurement science that will enable the identification of proven, scalable, and interoperable techniques and metrics to promote the development of AI.'[82] In July 2025, NIST released a draft outline for its first "Zero Draft" on AI Testing, Evaluation, Verification, and Validation (TEVV) for public feedback. This effort—part of NIST's AI Standards Zero Drafts pilot project—aims to accelerate the development of voluntary, consensus-based standards by creating stakeholder-informed proposals for formal standardization. In May 2024, NIST launched ARIA (Assessing Risks and Impacts of AI), a program to advance measurement science for safe and trustworthy AI. It aims to address gaps in AI evaluation that hinder generalization to real-world settings, improve understanding of AI's impacts on individuals and society, and provide organizations with critical information about the performance, reliability, and safety of AI systems once deployed.

Further investment is needed to enhance the impact of these and other promising initiatives and advance the science and practice of AI evaluation and testing. Ongoing public-private collaboration can accelerate progress, helping enable testing to serve as a coherent and effective tool within AI risk management and governance.

Engineering reliable foundations for AI evaluation and testing by strengthening rigor, standardization, and interpretability[83] will advance trust in AI, support broader adoption, and help actors across the AI supply chain meet emerging public policy and governance expectations. This requires focused effort across each of the three areas that are foundational to building a credible AI evaluation and testing ecosystem:

1. *Rigor* in defining what is being evaluated and why it matters as well as how an evaluation is being conducted. This requires either detailed specification of what is being measured and an understanding of how deployment context may affect outcomes—or transparency that an evaluation is exploratory in nature (as may be the case with, for example, open-ended red teaming).

[82] Pillar I, page 10.
[83] In this white paper, we use the term "interpretability" in a colloquial sense, referring to how understandable AI test results are to their intended audiences. This usage is distinct from the research field of mechanistic interpretability, which focuses on understanding the inner workings of frontier AI models.

2. *Standardization* of how evaluations and tests should be conducted to achieve valid, reliable results. This requires establishing technical standards that provide methodological guidance and promote quality and consistency. Where there is a need to account for context-specific risks, and specialized tools or methods are required, understanding what should be standardized and what requires bespoke consideration is also important. (Notably, this requirement for evaluations and tests with valid, reliable results is also distinct from what can be expected for exploratory research, which can continue to contribute complementary value with less standardization.)
3. *Interpretability* of evaluation and test results and clarification of how they inform risk decisions. This requires establishing expectations for evidence and improving literacy in how to understand, contextualize, and use results—while remaining aware of their limitations.

In the next section, we offer policy recommendations for advancing the AI evaluation and testing ecosystem, building from policy lessons from other domains and recognizing where multistakeholder AI initiatives are currently driving rapid progress.

Part 3: Discussion and policy recommendations

Policymakers and experts are actively working to prioritize the allocation of governance resources for assessing advanced AI models for at-scale public safety and national security risks. For instance, leading AI model developers' frontier safety policies—including Microsoft's Frontier Governance Framework—commit to conducting pre-deployment risk assessments that rely on extensive evaluations aimed at estimating the model's full capabilities, known as 'maximal capability evaluations'.[84] Developers also aim to evaluate the nature and scale of potentially harmful capabilities that remain after safety mitigations and guardrails have been applied, referred to as 'safeguard evaluations'.[85] These resource-intensive evaluations are undertaken to determine whether there are plausible pathways through which advanced AI models could be misused by malicious actors to develop biological weapons; significantly uplift cyberattacks; or accelerate the pace of AI development through the automation of expert-level research (among other potentially severe risks). As the AI Action Plan, the IASR, and the Singapore Consensus recognize, there is much work to be done in advancing frontier risk assessments and the model evaluations utilized to inform them.

However, AI models and end-to-end applications and services—including those that are not at the frontier of performance—may present a broader set of risks that become more visible in proximity to their real-world applications. The likelihood and impact of these risks depend heavily on how AI systems[86] are deployed in specific contexts, reflecting challenges seen with other GPTs like genome editing[87] and nanotechnology.[88] To measure and manage this broader set of risks, there is a need to advance sector-specific and context-aware evaluation frameworks, measurement instruments, and best practices. These should help post-deployment actors to effectively and efficiently identify, assess, and manage application-specific risks in the process of deploying AI systems. Recognizing this, NIST launched the ARIA program in 2024 to (among other aims) 'address gaps in AI evaluation that make it difficult to generalize AI functionality to the real world.'[89] Likewise, Singapore's AI Verify Foundation recently concluded its Global AI Assurance Pilot (Pilot), aiming to catalyze best practices for testing of generative AI applications. The Pilot found:

- Most current efforts in AI testing focus primarily on the safety of general-purpose AI models themselves, rather than on the reliability of complete, end-to-end AI applications and services.
- As generative AI moves from personal productivity tools and consumer-facing chatbots to deployment in physical and high-stakes environments—such as hospitals, airports, and banks—it faces stricter requirements for quality and reliability.

[84] Frontier Model Forum. (2024). *Issue Brief: Preliminary Taxonomy of Pre-Deployment Frontier AI Safety Evaluations*. https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-pre-deployment-frontier-ai-safety-evaluations/.
[85] Ibid.
[86] For the purposes of Part 3, we adopt the IASR definition of "system": '[a]n integrated setup that combines one or more AI models with other components, such as user interfaces or content filters, to produce an application that users can interact with.'
[87] Charo, A. and Greenfield, A. (2025). *Genome Editing Case Study*.
[88] Dionne, J. (2025). *Nanoscience and Nanotechnology Case Study*.
[89] NIST ARIA (Addressing Risks and Impacts of AI). (2024). AI Evaluations: Assessing Risks and Impacts of AI. https://ai-challenges.nist.gov/uassets/6.

- Integrating general-purpose AI models with existing data sources, workflows, and other system components adds complexity and greater variance in AI system behaviour. This creates more potential points of failure and challenges measurement.
- Risk assessments depend significantly on the context of the use case. For example, there is much lower tolerance for error in clinical applications compared to customer service chatbots.

Establishing robust foundations for AI evaluation and testing requires effort to improve rigor, standardization, and interpretability, ensuring that methods keep pace with rapid technological progress and evolving scientific understanding. Taking lessons from other GPTs such as genome editing and nanotechnology, as well as the recent findings of Singapore's Global AI Assurance Pilot, this foundational work must be pursued for both AI models and end-to-end AI applications and services. While testing models will continue to be important (especially for a narrow set of risks to public safety and national security), reliable evaluation tools that provide assurance for system performance downstream will enable broad and responsible adoption of AI. A strong feedback loop—on the design and implementation of evaluations for AI models and downstream applications and services, and between pre-deployment testing and post-deployment monitoring—could not only accelerate progress on methodological weaknesses but also bring focus to which opportunities, capabilities, risks, and impacts are most appropriate and efficient to evaluate at what points along the AI development and deployment lifecycle. This clarity not only saves resources but also supports prioritization of risk management efforts where they are best applied.

Based on this context, we offer the following six policy recommendations to strengthen AI evaluation and testing as part of effective AI governance:

1. **Bring greater focus to advancing testing at the level of end-to-end AI applications and services while also continuing to advance testing at the model level.**

While many AI policy frameworks establish an expectation for testing end-to-end AI applications and services, much investment in developing evaluation methodologies and tools has been focused on evaluating AI models, as Singapore's AI Verify Pilot recently found. Recognizing that AI is a GPT that presents a wide range of opportunities, capabilities, risks, and impacts across deployment scenarios, there is an urgent need to advance AI evaluation and testing best practices and tools for those building and deploying end-to-end AI applications and services, especially in high-stakes deployment environments such as hospitals. Work on advancing evaluation science and practical measurement tools for AI models and end-to-end applications and services would best be pursued in close coordination. This could enable insights gained on methodologies to have greater impact and help direct limited resources to the opportunities, capabilities, risks, and impacts most effective and efficient to evaluate and test at each level.

2. **Continue to invest in multistakeholder collaboration to strengthen the scientific foundations, methodological rigor, and infrastructure needed for rigorous evaluation.**

There is strong recognition of the need to strengthen the scientific foundations of

AI evaluation.[90] Much work is needed not only to develop clear and systematic definitions of capabilities, risks, impacts, measurement objectives, and thresholds for AI performance but also to formulate consensus standards for AI evaluation and risk assessment methods. Especially in the context of advanced general-purpose AI, policymakers should encourage analyses that go beyond what is easy to measure, exploring interdependencies between capabilities, risks, and impacts that are often currently treated in siloed categories. To do this, efforts need to incorporate multistakeholder collaboration, recognizing the many ongoing initiatives surveyed in Part 2 and the value of cumulative and coherent progress.

3. **Leverage learnings across sector-specific evaluation frameworks that are tailored to high-risk settings[91] for AI systems and can be operationalized by a broad range of actors.**

In many cases, risk assessment frameworks already exist in high-risk sectors but need to be adapted for generative AI. This work should be accompanied by investment in the development of measurement instruments, practical tools, and training programs to uplift the capabilities of AI evaluators and post-deployment actors so they are effective evaluators and consumers of evaluation results. As progress is made in sector-specific evaluation frameworks, learnings can be shared and built upon as relevant in other sectors, offering some of the efficiencies sought through horizontal governance approaches.

4. **Strengthen partnerships between governments, experts, and industry to provide best practice guidance and build norms to support AI evaluation and risk mitigation across the AI lifecycle.**

For effective cross-lifecycle AI governance, information generated through post-deployment monitoring is a critical input. So too are ongoing risk assessments and feedback loops that inform and improve the efficiency and effectiveness of pre-deployment evaluations. One important piece of this puzzle is the development of robust infrastructures for identifying and addressing flaws in deployed systems. Experts—including those from MIT, Stanford, Princeton, MLCommons, and Microsoft—have observed that the infrastructure, practices, and norms for reporting flaws in general-purpose AI systems remain seriously underdeveloped, 'lagging far behind more established fields like software security.'[92] Building norms for reporting, addressing, and disclosing flaws or patterns of misuse post-deployment can enhance cross-ecosystem understanding of risk and implementation of mitigations.

5. **Weigh trade-offs with emphasizing any particular aspect of an AI evaluation framework, recognizing the influence of early governance decisions.**

Choices about what to emphasize in an AI evaluation framework—such as pre-deployment versus post-deployment testing, internal versus external assessments, or evaluation for regulatory enforcement versus risk management—shape how the ecosystem develops over time. Experience from other domains shows that these early decisions can become

[90] IASR, section 3.3; Singapore Consensus (2025), section 1; this paper's Introduction.
[91] Numerous frameworks classify AI systems deployment settings as involving higher risk. For example, the EU AI Act designates certain AI systems used in settings such as education, employment, law enforcement, and migration as high-risk (see Annex III), and the US has designated certain systems "high-impact" AI for federal agencies (see Section 6, Purposes for Which AI Is Presumed to be High Impact, in Memorandum No. M-25-21). Microsoft takes a similar approach in its internal AI governance program, providing hands-on counseling for higher-risk AI use cases through the Sensitive Uses and Emerging Technologies team.
[92] Longpre, S., et al. (2025). *In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI*. https://arxiv.org/abs/2503.16861.

entrenched, even when challenges emerge. Given the rapid evolution of AI capabilities and governance needs, it is critical to be deliberate about where attention and resources are directed. Iteration is more difficult in practice than in principle, and misaligned emphasis can divert risk management efforts and have lasting impacts on governance objectives.

6. **Deepen attention to the role of transparency as a foundation for shared understanding and iterative improvement in AI evaluation and governance.**

While transparency is often cited as a core principle of AI governance, its role in supporting effective evaluation and risk management warrants deeper reflection. Different types of transparency—ranging from documentation and disclosures to evaluation artifacts—can serve distinct purposes across the AI lifecycle. For example, appropriate forms of transparency can help contextualize evaluation results, making them more interpretable and actionable in deployment settings. Transparency frameworks—when designed to appropriately balance national security and intellectual property disclosure risks—enable shared learning and adaptive feedback loops.[93] They provide a critical foundation for iterative improvement in evaluation practices and broader risk management strategies, especially as methodologies, capabilities, and risks continue to evolve. Greater deliberation and clarity are needed, however, on what kinds of transparency are most appropriate for which actors, at which stages of the AI lifecycle, and to what ends.

[93] Bommasani, R., et al. (2025). *The California Report on Frontier AI Policy*. https://www.arxiv.org/abs/2506.17303.

Appendix A: Types of AI evaluation and testing

The International AI Safety Report (2025) (IASR) defines evaluations as '[s]ystematic assessments of an AI system's performance, capabilities, vulnerabilities or potential impacts. Evaluations can include benchmarking, red-teaming and audits and can be conducted both before and after model deployment.'[94] Reflecting the nascency of practices and ongoing experimentation by evaluators, there is not yet a widely agreed typology of AI evaluation methods though different, sometimes contradictory, definitions have emerged. Broadly, types differ based on their object (e.g., model, system or application); timing (e.g., pre-deployment or post-deployment); what, conceptually, is being measured (e.g., whether the evaluation is to make claims about an attribute, behavior or impact of the system); the method of sourcing the dataset for evaluation (e.g., human-created input test cases, sampling real traffic, or utilizing synthetic data); and other factors. We non-exhaustively define types in the table below, drawing on the IASR, guidance published by the Frontier Model Forum, and input from internal and external experts. We also direct readers to:

- NIST's Assessing Risks and Impacts of AI (ARIA) Program Evaluation Design Document (2024) for terminology relevant to evaluation of AI applications.
- For a more detailed articulation of frontier AI model evaluations, the Frontier Model Forum's Preliminary Taxonomy of Pre-Deployment Frontier AI Safety Evaluations (2024) and Technical Report on Frontier Capability Assessments (2025).

Auditing	A formal review of an organization's compliance with standards, policies, and procedures, typically carried out by an independent third party.[95]
Benchmark testing	A benchmark is a standardized, often quantitative test or metric used to evaluate and compare the performance of AI models or systems on a fixed set of tasks (such as multiple-choice questions) designed to represent real-world usage.[96] Benchmark testing aims to use these standardized criteria to quantify the performance of AI models or systems in such a way that results can be compared at scale, over time, and across models or systems.[97]
AI red teaming	A systematic process in which dedicated individuals, teams, or tools search for vulnerabilities, limitations, or potential for misuse through various methods.[98] Often, the red team searches for inputs that induce undesirable behaviour in a model or system to identify safety gaps. Microsoft's AI Red Team explains, 'AI red teaming strives to push beyond model-level safety benchmarks by emulating real-world attacks against end-to-end systems.'[99]

[94] IASR, Glossary.
[95] Ibid.
[96] Ibid.
[97] Frontier Model Forum. (2024). *Issue Brief: Preliminary Taxonomy of Pre-Deployment Frontier AI Safety Evaluations.* https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-pre-deployment-frontier-ai-safety-evaluations/.
[98] IASR, Glossary.
[99] Bullwinkel, B., et al. (2025). Lessons From Red Teaming 100 Generative AI Products. https://arxiv.org/abs/2501.07238.

Uplift studies	Assessing how advanced AI systems might be used by malicious actors to carry out real-life harmful tasks, compared to the use of existing tools such as internet search.[100] Such studies typically utilize a controlled trial with a treatment group that has access to AI and a baseline or control group that is limited to alternate resources. The aim of these rigorous studies is to approach a grounded assessment of the counterfactual impact of AI on the capabilities of malicious actors.
Risk or safety-focused evaluations	Targeted context-aware evaluations of risk (i.e., probability of an event x severity of its impacts) as opposed to an AI model's or system's isolated attributes. While risk-specific benchmarks are being developed, holistic evaluations require examination of both event and impact, often requiring mixed measurement methods (because impacts cannot be observed in a sandbox). Sometimes, proxies for impact can be employed to simplify the task.
Policy adherence evaluations[101]	Targeted evaluations of whether an AI system's in-context behaviors align with a set of policy requirements. This typically involves monitoring an AI system's output for certain behaviors and phenomena that are forbidden by the reference policy.

[100] Frontier Model Forum. (2024). *Issue Brief: Preliminary Taxonomy of Pre-Deployment Frontier AI Safety Evaluations*. https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-pre-deployment-frontier-ai-safety-evaluations/.
[101] This is not a broadly utilized term but rather a descriptive one.

Appendix B: Case summaries

This summary table is not a substitute for the nuanced insights available in each standalone expert case study. It provides high-level comparisons to complement those insights.

Domain	What to test for & setting thresholds	How to test	When to test	Role within governance
Genome Editing *Use of genome editing to modify or enhance varied products (e.g., foods and drugs).*	Intended and unintended off-target genetic effects and phenotypic effects. No general thresholds. Dependent on application context (e.g., more qualitative in agriculture and more quantitative in pharmaceuticals).	Testing and risk assessments are sector-specific based on the context in which genome editing is applied.	Sector-specific.	Sector-specific.
Nanoscience and Nanotechnology *Nanomaterials or varied products (e.g., semiconductors and cosmetics) incorporating nanomaterials.*	In laboratory, testing focused on safety and effectiveness of nanomaterials across synthesis, deployment and disposal. Thresholds vary across sectors, application contexts, and jurisdictions.	Multifaceted testing landscape that is sector specific.	Sector-specific.	Sector-specific.
Financial Services (Bank Stress Testing) *Capital adequacy of banks under predicted conditions of stress.*	Resilience of banks under macroeconomic stresses including capital adequacy and liquidity. Adverse scenarios (approximating recessionary conditions) define capital adequacy thresholds.	Stress testing: regulator provides adverse scenarios, banks provide data, and regulator uses confidential model to predict how bank will perform under stress. Exploratory analysis emerging.	In many jurisdictions including the US, stress testing is an annual exercise.	Banks of a certain size or significance are required to participate. Stress testing is used as a regulatory tool for transparency, public supervision and oversight, and to better understand dynamic risks.
Cybersecurity *Security of digital systems and networks.*	Vulnerability to adversarial attacks and resilience to various security breaches. Qualitative thresholds relying on light-tough security standards and industry self-assessments.	Multiple and evolving methods including penetration testing, red-teaming, vulnerability assessments, continuous monitoring and incident analysis.	Cybersecurity testing should be ongoing. Institutions are encouraged to conduct regular penetration tests and maintain monitoring capacities.	Important but flexile role in governance, with speed of technological change challenging highly prescriptive rules. Critical infrastructure sectors legislatively required to meet best practices.
Pharmaceuticals *Pharmaceutical compound or drug candidate.*	Benefits, safety and efficacy. Generally defined regulatory thresholds with more disease-specific guidelines issued by FDA due to variance across disease types.	Multiple methods including preclinical lab tests, phased clinical trials (Phases I-III), and post-market surveillance (Phase IV).	Throughout lifecycle: pre-market (Phases I-III) and post-market (surveillance and Phase IV). In practice, post-market monitoring is limited.	Testing is mandatory and central to regulatory approvals and market access.
Medical Devices *Medical devices including hardware and software.*	Safety, performance usability, quality (e.g., failure rates, side effects). Quantitative, risk-tiered thresholds tied to device class and intended use; higher classes require clinical evidence.	Multiple methods including bench testing, software validation, clinical trials, post-market surveillance and conformity assessments.	Throughout the lifecycle: pre-market (e.g., design validation) and post-market (real-world surveillance and tracking of devices).	Testing is mandatory and linked to risk-based classifications and demonstrating regulatory compliance. De novo process for new innovations.

Civil Aviation *Aircraft and aircraft systems (e.g., engines, avionics etc.)*	Quantitative thresholds for airworthiness of designs and manufactured products defined in extremely detailed technical standards and FAA certification rules. Case-by-case definition in certain contexts.	Multiple methods including simulation and modelling, flight testing, and engineering analysis.	Throughout the lifecycle: design, construction, operation, certification and when the aircraft is in-service.	Testing is mandatory for airworthiness certification and tightly regulated.
Nuclear Power *Nuclear facilities and structures, systems and componentry (SSCs).*	System safety, functionality, environmental and radiological risk, durability. Quantitative thresholds based on risk probabilities and health consequences; often codified in national regulations.	Multiple methods including benchmarking of SSCs, prototype testing, periodic inspections, and environmental monitoring.	Throughout the lifecycle: from design and construction to operation and decommissioning of nuclear facilities.	Testing is mandatory and embedded in licensing and safety regimes.