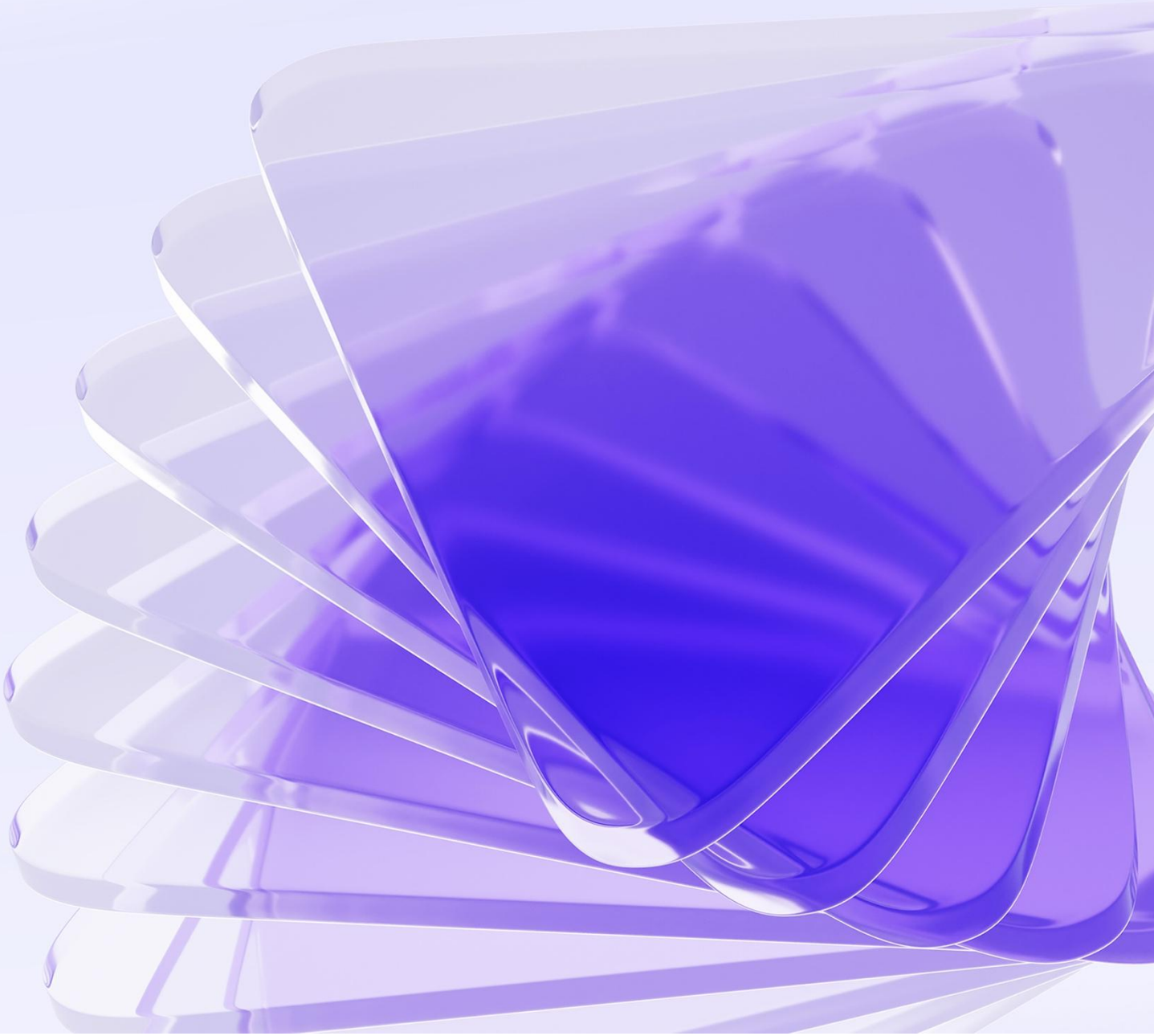
NPL
National Physical Laboratory

From Frameworks to Protocols:

A Measurement Science Foundation for Technical AI Assurance



National Physical Laboratory

National Physical Laboratory (NPL)

From Frameworks to Protocols:

A Measurement Science Foundation for Technical AI Assurance

Chris P. Barnes, James Adams, Spencer A. Thomas, Paul Duncan

We are the UK's National Metrology Institute (NMI), a world-leading centre of excellence that provides cutting-edge measurement science, engineering and technology that underpins prosperity and quality of life in the UK.

Contents

Executive summary 3

Glossary 5

Introduction 7

Landscape of existing AI frameworks 13

The missing measurement layer 26

Challenges for AI measurement 29

Protocols as a foundation layer for technical AI assurance 32

A proposed protocol library for the measurement of AI systems 35

From protocols to technical standards 40

Next steps for the Centre for AI Measurement 41

Summary 42

Executive summary

As artificial intelligence (AI) systems become more capable and integrated into critical workflows, the need for robust assurance mechanisms continues to grow. These mechanisms are essential for providing confidence in AI-enabled decisions and for demonstrating the quality, reliability and fitness-for-purpose of AI systems. Over the last decade, substantial progress has been made in the development of AI governance frameworks, risk management approaches and emerging regulatory regimes. The UK has made significant investments in AI safety, assurance, standards and quality infrastructure, creating an opportunity to establish international leadership in technical AI assurance. However, a significant gap remains between high-level governance objectives and the practical generation of technical evidence needed to support assurance decisions.

This report provides a non-exhaustive discussion of the various frameworks currently in use. It then argues that AI assurance requires a stronger foundation in measurement science. While existing frameworks are effective at defining what trustworthy AI should achieve, comparatively little attention has been paid to how trustworthiness properties should be measured. These challenges are amplified by the complexity of modern AI systems, including foundation models, retrieval-augmented generation systems and agentic AI, whose behaviour is often stochastic, context-dependent and difficult to characterise using conventional evaluation methods.

To address this gap, the report proposes a protocol-based approach to technical AI assurance. Through the Centre for AI Measurement, the National Physical Laboratory (NPL) is developing these protocol approaches as part of a wider effort to establish the measurement foundations required for trustworthy AI. Protocols are positioned as a missing intermediate layer between governance frameworks and formal standards, providing structured, repeatable, auditable and interoperable methods for generating assurance evidence. Drawing on concepts from metrology, quality infrastructure, and credibility and conformity assessment, protocols establish common approaches for defining measurands, conducting evaluations, quantifying uncertainty, ensuring traceability and documenting results. Unlike fixed standards, protocols can evolve rapidly in response to technological advances while retaining sufficient rigour to support reproducible assurance activities. Protocols can be used to contribute to standardisation efforts

underway and form the basis of future standards. By combining measurement science principles with protocol-driven evaluation, the report outlines a pathway from governance principles to standards-ready assurance methods capable of supporting trustworthy, vendor-agnostic AI deployment at scale and ultimately stimulating a competitive assurance ecosystem.

Glossary

AI Assurance: The process of building justified confidence that an AI system is safe, reliable, secure, effective and fit for its intended purpose through the systematic collection and evaluation of evidence.

Conformity Assessment: Demonstration that specified requirements relating to a product, process, system, person or body are fulfilled.

Evaluation: The assessment of the degree to which an AI system exhibits a defined characteristic, such as robustness, fairness or accuracy.

Framework: A structured set of principles, categories and processes used to govern, assess and manage AI systems and associated risks throughout their lifecycle.

National Metrology Institute (NMI): The organisation designated to realise, maintain and disseminate national measurement standards and provide the highest level of metrological traceability within a country.

Measurement: The process of experimentally obtaining one or more values that can reasonably be attributed to a quantity together with any other available relevant information.

Measurand: The quantity intended to be measured. In some cases, the measurand is specified additionally by a documented and accepted measurement procedure.

Metrology: The science of measurement, providing the principles and methods needed to ensure measurements are reliable, reproducible, comparable and fit for purpose.

Protocol: A complete, repeatable and pre-declared procedure for generating technical evidence about one coherent property of a specified AI system under defined conditions.

Quality Infrastructure: The ecosystem of standards, metrology, accreditation and conformity assessment organisations that collectively provide confidence in technical claims and measurements.

Testing: Activities that determine whether a system, component or work product meets specified requirements or expectations.

TEVV (Testing, Evaluation, Verification and Validation): A collection of technical activities used to generate evidence regarding the performance, safety, reliability and trustworthiness of AI systems.

Traceability: The ability to relate a measurement or evaluation result to the data, models, tools, configurations and procedures used to generate it.

Uncertainty: A parameter, associated with the result of a measurement, that characterises the dispersion of the values that could reasonably be attributed to the measurand.

Validation: Confirmation that a system is fit for its intended purpose within a specified operational context.

Verification: Confirmation, through objective evidence, that specified requirements have been fulfilled and that the system has been built correctly.

Introduction

Artificial intelligence (AI) is increasingly being deployed in high-impact domains including healthcare, finance, defence and public services, where failures can have significant economic, societal and safety consequences. As the use cases become more complex and less transparent, the risks increase and the assurance mechanisms need to be proportionately more robust. AI assurance has emerged as a mechanism for generating evidence that systems are trustworthy, safe, reliable, secure and fit for their intended purpose, thereby increasing confidence in both the quality of AI systems and the decisions they support [1]. Assurance activities help organisations evaluate whether AI systems operate as expected within defined conditions, identify limitations and risks, and communicate these findings to a broad body of stakeholders such as users, regulators, insurers and procurement bodies. Additionally, assurance should demonstrate the benefits of deploying AI systems [2].

The importance of assurance is reflected in growing international attention from governments, standards bodies and regulators, including the UK government's Artificial Intelligence Playbook [3], the NIST AI Risk Management Framework [4], the OECD AI Principles [5], the EU Product Liability Directive [6], the EU AI Act [7], the capAI procedure [8] and ISO/IEC standards development [9]. Collectively, these initiatives recognise that trust in AI depends not only on innovation but also on the availability of robust evidence demonstrating that systems perform safely and responsibly throughout their lifecycle. A lack of trust and confidence in AI systems is the primary barrier to large-scale adoption in complex and high-risk settings such as industry [10].

Despite increasing recognition of its importance, widespread adoption of AI assurance remains challenging due to key market failures. For instance, organisations face technical barriers in both deployment and assurance, including limited transparency over training data and model development, difficulties accessing and deploying proprietary systems, a lack of agreed metrics and benchmarks, poor treatment of uncertainty and limited evidence that laboratory evaluations predict real-world performance. Furthermore, assurance practices are often bespoke, expert-driven and difficult to compare across providers, creating challenges for procurement, certification, affordability and regulatory oversight. In addition, current assurance practices are not dynamic enough to keep up with rapid development of AI systems [11]. The UK government report ‘Assuring a Responsible Future for AI’, indicated that public confidence in AI systems cannot rely solely on claims made by developers and that independent assessment and assurance are important mechanisms for building trust, particularly for high-impact and safety-critical applications [12].

To address these issues, the UK government has progressively invested in developing an AI assurance ecosystem and, in 2025, published the 'Trusted Third-Party AI Assurance Roadmap', which identifies four key barriers to market growth: assurance quality, skills shortages, limited access to information and insufficient innovation [13]. The roadmap positions independent third-party assurance as a critical enabler of trustworthy AI adoption, providing mechanisms for the objective evaluation of AI systems and helping to build confidence among developers, deployers, regulators and the public. As a result, the AI Assurance Innovation Fund was introduced to accelerate the development of AI assurance capabilities through the Centre for AI Measurement at NPL [14], and the UK AI Assurance Stakeholder consortium created to focus on delivering a skills framework and profession supporting AI assurance. In parallel, the UK government's Digital Standards Strategy 2026-2030 also identifies standards as a critical mechanism for supporting innovation, interoperability, trust and international competitiveness in emerging digital technologies, including AI [15]. The strategy emphasises the role of standards not only in supporting regulation but also in accelerating adoption, reducing market fragmentation and enabling confidence in new technologies.

In this report we highlight some of the many initiatives relevant to the growing AI assurance ecosystem. These capture a snapshot of the evolving landscape and are used for illustrative purposes. Due to the rapid and international nature of AI assurance development, capturing every activity would be impossible. However, what we argue is that while governance frameworks and assurance guidance continue to mature, a significant gap remains between high-level principles and the technical practices required to generate reliable evidence [16]. Current frameworks generally articulate what characteristics trustworthy AI should possess but provide limited guidance on how such properties should be measured, quantified and compared, let alone the tools to obtain them. Consequently, assurance remains fragmented, with substantial variation in evaluation methodologies, reporting practices and evidential standards.

This report argues that a measurement science mission, through the Centre for AI Measurement, can help address the AI assurance challenge by introducing principles of traceability, reproducibility, uncertainty quantification and comparability into AI measurement. Through the development of structured and modular assurance protocols, measurement science offers a pathway towards more robust, scalable and standards-ready approaches capable of supporting the assurance needs of contemporary and emerging AI systems, including foundation models and agentic AI. Ultimately this enables innovation and market adoption; by providing credible, objective evidence of safety, performance and compliance, assurance can help organisations bring new AI-enabled products and services to market more quickly and with greater confidence. Common assurance approaches and evaluation protocols also create a more level playing field across the

ecosystem, helping SMEs and larger technology companies alike demonstrate trustworthiness against shared expectations. In this way, assurance supports both responsible innovation and economic growth, while fulfilling a key public-interest role in providing confidence to government, industry and society.

Definitions

To avoid ambiguity, this report distinguishes between governance frameworks, technical assurance activities, and the broader assurance process. While these terms are often used interchangeably in the literature, they represent different levels of abstraction and serve distinct purposes within the AI assurance ecosystem.

Frameworks define a structured set of principles, categories and processes that helps organisations identify, assess, manage and mitigate risks associated with AI systems throughout their lifecycle.

They provide guidance on what should be considered when governing AI systems, including issues such as safety, robustness, fairness, accountability and transparency. However, they typically do not prescribe the specific methods or measurement procedures required to generate evidence that these properties have been achieved.

Technical AI Assurance is the application of defined processes, methods and measurement techniques to generate objective and reproducible evidence regarding the performance, safety, reliability and trustworthiness of an AI system under stated conditions.

Technical assurance is concerned with the collation, generation, interpretation and communication of evidence. This evidence can support claims about system behaviour and suitability for deployment but also answer queries from a regulator. In contrast to governance frameworks, technical assurance focuses on how evidence is obtained, evaluated and reported. It therefore encompasses activities such as testing, evaluation, verification, validation (TEVV) and ongoing monitoring.

Testing is a set of activities that determine whether a work product meets specified requirements or expectations. [9]

The output of a test is normally a binary outcome: pass or fail. This outcome may be based on a direct observation or a measurement that characterises a particular system property.

Evaluation is a set of activities that determine the degree to which a work product possesses a defined characteristic. [9]

Evaluation determines the significance of measurement results and whether they support a particular claim about system performance or behaviour. From a measurement science perspective, this assessment naturally requires the measurement result to include an associated uncertainty.

Verification is confirmation, through the provision of objective evidence, that specified requirements have been fulfilled. [17]

Verification addresses the question "Was the system built correctly?" It assesses whether a system conforms to stated specifications, requirements, design intentions or relevant standards. In this report, verification refers to evidence-based confirmation through testing and assessment and could include, but is not limited to, formal verification methods that seek mathematical proofs of correctness.

Validation is confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled. [18]

Validation addresses the question "Are we using the right system?" It assesses whether a system is fit for its intended purpose within a specified operational context and whether it satisfies stakeholder needs in practice.

Assurance is the broader process of building justified confidence in a system through the systematic collection, evaluation and communication of evidence.

Assurance integrates outputs from testing, evaluation, verification, validation and other supporting activities to determine whether a system meets specified requirements and is acceptable for use in each application. As such, assurance is not a single activity, but an evidence-based judgement supported by an auditable body of evidence. A useful distinction is that TEVV activities generate evidence, whereas assurance uses that evidence to support confidence and decision-making. Testing may produce measurements, evaluation may interpret those measurements, and verification and validation may establish compliance and fitness for purpose, but assurance is the process that brings these elements together to justify trust in the system. This distinction is central to the measurement science perspective adopted throughout this report.

Quality infrastructure

In established technical domains, confidence in measurements and conformity assessments is supported by a broader quality infrastructure. Quality infrastructure comprises the organisations that collectively provide confidence in the quality, reliability and comparability of technical claims and products. These include regulators, accreditation bodies, government, standards bodies, research bodies, national measurement capability, civil society organisations and professional bodies. National quality infrastructures have been fundamental to sectors such as healthcare, manufacturing, telecommunications, energy and environmental monitoring, where measurement results must be accepted across organisations, sectors and national boundaries.

The same principles are increasingly relevant for AI. Within the UK, several organisations have particularly important roles in the development of AI quality infrastructure. The British Standards Institution (BSI) plays a central role in standards development, representing UK interests in international standards activities and contributing to AI-related standards within ISO/IEC and CEN-CENELEC. BSI has also been a key participant in the UK's wider AI assurance ecosystem and is among the first certification bodies accredited to certify organisations against ISO/IEC 42001, an international standard for AI management systems. The United Kingdom Accreditation Service (UKAS) provides independent accreditation of conformity assessment organisations, including testing laboratories, inspection bodies and certification organisations. Accreditation provides confidence that assessments are performed competently, consistently and impartially, helping to ensure the quality and credibility of assurance results. As AI assurance matures, UKAS is likely to play an analogous role to that seen in established technical sectors, enabling confidence that assurance providers are applying approved protocols and methodologies correctly. Independent accreditation may ultimately become an important mechanism for ensuring the credibility and comparability of AI assurance results across assurance providers and sectors.

The NPL, as the UK's National Metrology Institute (NMI), occupies a complementary position within the UK's quality infrastructure. The traditional NMI role is not regulation or certification, but the development of measurement science, reference methods, uncertainty analysis and inter-comparison methodologies. As a government-owned body it is independent and removed from market forces and focused on delivering impact for industry and the UK. In the context of AI assurance, the recent establishment of the Centre for AI Measurement solidifies NPL's role in developing and validating assurance protocols, establishing best practice for measurement and uncertainty treatment, supporting inter-laboratory comparisons, and contributing technical expertise to standards development [19]. Such activities help ensure that assurance evidence remains scientifically robust and reproducible as assurance practices mature. The AI Security

Institute (AISI) is another important player within the UK's emerging AI assurance ecosystem, providing independent evaluation and research capabilities focused on frontier AI systems. Through the development of advanced evaluation methodologies, benchmarks and red-teaming approaches, AISI is helping establish the scientific foundations for assessing AI capabilities, risks and safety properties [20].

The AI Quality Infrastructure (AIQI) Consortium was established to bring together organisations from across the international quality infrastructure ecosystem and promote a coordinated approach to the safe, secure and ethical development of AI. The consortium emphasises the role of consensus-based standards, accredited testing and certification, and independent conformity assessment as essential mechanisms for building trust in AI systems [21]. Within the AIQI ecosystem, standards organisations develop technical requirements and guidance; metrology institutions provide the measurement foundations needed to ensure reliability and comparability; testing, inspection and certification bodies assess conformance against defined requirements; and accreditation bodies provide independent confirmation that these activities are undertaken competently and impartially. The AIQI Consortium argues that a similar combination of metrology, standards, conformity assessment and accreditation will be required if AI systems are to be trusted by regulators, industry and the public.

Landscape of existing AI frameworks

Over the past decade, a diverse ecosystem of AI governance, risk management and assurance frameworks has emerged in response to growing concerns around the safety, reliability, fairness and accountability of AI systems. These frameworks have played an important role in establishing common language, defining desirable characteristics of trustworthy AI, and supporting organisations in identifying and managing AI-related risks. However, they vary significantly in scope, purpose and level of technical detail and often stop short of specifying the rigorous and reproducible evaluation procedures required to support independent assurance activities. Table 1 gives a summary of a selection of frameworks discussed below.

Framework	Organisation	Primary Purpose	Scope	Type
OECD AI Principles	OECD	Policy principles	International	Principles
Recommendation on the Ethics of Artificial Intelligence	UNESCO	Policy principles	International	Principles
NIST AI RMF	NIST	Risk management	Organisation-wide	Governance & risk management
NPL Trustworthy & Safe AI Lifecycle	NPL	Lifecycle governance and assurance	End-to-end AI lifecycle	Governance & risk management
SAFE AI Framework	CDAC Network and Alan Turing Institute	Governance for design and deployment of AI systems	Organisation-wide AI governance across the AI lifecycle	Governance & risk management
ISO/IEC 42001	ISO/IEC	AI management system	Organisation-wide	Governance & risk management
ISO/IEC 23894	ISO/IEC	AI risk management	Organisation-wide	Governance & risk management
AI Management Essentials (AIME)	UK Government	AI management and governance	Organisation-wide	Governance & risk management
AI Risk Management Toolkit	UK Government	AI risk management	Full lifecycle	Governance & risk management
EU AI Act	European Union	Regulation and compliance	High-risk AI systems	Regulation, policy & compliance

AMLAS	University of York	Safety assurance	ML-enabled autonomous systems	Assurance methodologies
SACE	University of York	Safety assurance	Autonomous systems	Assurance methodologies
AI Verify Testing Framework	Singapore IMDA	Assurance and testing	Conventional and Generative AI	Assurance methodologies
SAIF	Google	AI security assurance	AI development lifecycle	Assurance methodologies
The Cross-Government AI Testing and Assurance Framework	UK Cross-Government Testing Community	AI testing and assurance guide	Full life cycle	Assurance methodologies
NIST Zero Draft	NIST	Accelerate standardisation	Broad applicability based around community input	Evaluation standard

Table 1: Frameworks for AI governance and assurance. Rows are coloured by type of framework

Principles

At the highest level of the AI governance landscape are principles-based frameworks, which seek to define the characteristics of trustworthy and responsible AI rather than prescribe specific implementation mechanisms. These frameworks have played a critical role in establishing international consensus around desirable AI behaviours and have influenced many subsequent standards, regulations and assurance approaches. The most influential example is the OECD AI Principles, which were adopted in 2019 and have subsequently informed AI policy across numerous jurisdictions [5]. The principles promote inclusive growth, human-centred values, transparency, robustness, security and accountability, while encouraging organisations to manage AI systems throughout their lifecycle. Their significance lies not in providing operational guidance, but in creating a common language and shared understanding of what trustworthy AI should look like. The UNESCO Recommendation on the Ethics of Artificial Intelligence builds upon and complements the OECD AI Principles. While both promote trustworthy, human-centred AI through principles such as transparency, accountability, fairness, and human oversight, UNESCO places greater emphasis on ethical, social, cultural, and environmental impacts and provides more detailed guidance for policy implementation [22].

Governance and risk management

Sitting between high-level principles and operational assurance activities, governance and risk management frameworks provide organisations with structured approaches for managing AI-related risks throughout the system lifecycle [23]. A prominent example is the NIST AI Risk Management Framework (AI RMF), which provides a voluntary, risk-based approach for governing AI systems throughout their lifecycle [4]. The framework is structured around four core functions: Govern, Map, Measure and Manage, encouraging organisations to identify risks, assess impacts and implement appropriate controls. A key strength of the AI RMF is its flexibility and broad applicability across sectors and use cases. It has become highly influential internationally and has helped establish a shared vocabulary for AI risk management. The associated Playbook [24] is a practical companion to the AI RMF providing detailed guidance, examples and suggested actions for managing AI risks, while acknowledging the importance of measurement and evaluation. However, it provides relatively limited guidance on the specific methodologies, evidence requirements and uncertainty treatments needed to demonstrate that a system satisfies particular trustworthiness requirements. A step in this direction was developed in the NPL Trustworthy & Safe AI Lifecycle [25], which builds on the AI RMF by describing how to map risks to measurable properties. The framework therefore begins to bridge governance and technical assurance, although many of the underlying measurement procedures, performance criteria and reporting mechanisms remained open to ongoing research and development. The SAFE-D principles (Sustainability, Accountability, Fairness, Explainability, Data Stewardship) provide another risk-based framework designed to support reasoning around the harms and benefits of AI systems and other data-driven technologies, particularly in sensitive domains such as healthcare [26]. In practice, SAFE-D acts as a set of high-level ethical goals that can be translated into concrete requirements, evidence, and assurance arguments throughout the lifecycle of an AI system, helping developers, regulators and users build justified confidence in its deployment. Nevertheless, like many governance frameworks, SAFE-D primarily describes desirable outcomes and organisational controls rather than providing standardised methods for measuring and validating system behaviour.

More recently, standards on AI governance and management have emerged. Published in 2023, ISO/IEC 42001 is the first international standard specifying requirements for an Artificial Intelligence Management System (AIMS) [27]. The standard follows the management-system approach used in standards such as ISO 9001 (quality management) [28] and ISO 27001 (information security) [29], providing organisations with a structured framework for governing the development, deployment and use of AI systems. Its primary objective is to help organisations establish policies, processes, responsibilities and controls that support the responsible use of AI

while balancing innovation, risk management and regulatory compliance. The standard is applicable to organisations of any size that either develop or use AI-enabled products and services. ISO/IEC 23894 complements ISO/IEC 42001 by providing guidance on the identification, assessment, treatment and monitoring of AI-specific risks [30]. Whereas ISO/IEC 42001 establishes the organisational management system, ISO/IEC 23894 focuses more specifically on risk management activities and can be viewed as the AI counterpart to established enterprise risk management approaches. The standard recognises that AI systems introduce distinctive challenges, including uncertainty, emergent behaviour, model drift, data quality issues, bias, security vulnerabilities and impacts on individuals and society. It therefore extends traditional risk management principles to account for the unique characteristics of AI technologies. In summary, governance frameworks provide increasingly mature approaches for managing AI risks, but generally stop short of defining the detailed measurement and evaluation methods needed for independent assurance.

The UK Government has also developed the AI Management Essentials (AIME) tool to support organisations in establishing baseline AI management practices [31]. AIME is a self-assessment framework designed primarily for SMEs and organisations seeking a practical entry point into AI governance. Drawing on principles from ISO/IEC 42001, the NIST AI Risk Management Framework and the EU AI Act [7], it helps organisations evaluate the maturity of their AI management systems and identify areas for improvement. More recently, the UK Government published the AI Risk Management Toolkit, which provides practical guidance for identifying, assessing, treating, monitoring and reporting risks associated with AI systems throughout their lifecycle [32]. The toolkit is intended to support teams across the design, procurement, deployment and operation of AI-enabled products and services, aligning AI risk management with established government risk-management practices. Similar to ISO/IEC 23894 and the NIST AI RMF, the toolkit focuses on governance processes and decision making rather than prescribing specific testing methodologies or assurance procedures. As such, it provides a structured approach for managing AI risks while leaving open the question of how evidence relating to those risks should be generated and assessed.

Regulation, policy and compliance

A distinct layer within the AI governance ecosystem is formed by regulatory, policy and compliance frameworks. These mechanisms translate broad governance objectives into practical obligations, expectations and incentives that influence how AI systems are developed, deployed and assured. Unlike principles-based and risk-management frameworks, which are generally voluntary, regulatory approaches create varying degrees of accountability for organisations using or

developing AI systems. From the perspective of technical AI assurance, this layer is important because it creates demand for evidence. Whether through regulation, standards, procurement requirements, certification schemes or sector-specific guidance, organisations increasingly need objective evidence demonstrating that AI systems are safe, reliable and fit for their intended purpose.

Various jurisdictions have adopted significantly different approaches. The most prominent legislative example is the European Union Artificial Intelligence Act, which introduces a risk-based framework with legally binding obligations for providers and deployers of AI systems [7]. The Act establishes requirements relating to risk management, data governance, robustness, transparency, human oversight and post-market monitoring, creating a clear demand for assurance evidence throughout the system lifecycle. However, while the Act defines what organisations must achieve, many of the detailed testing, evaluation and conformity assessment methods are still being developed through the CEN-CENELEC JTC 21 harmonised standards programme [33], [34]. Consequently, translating regulatory requirements into practical assurance activities remains an active area of standardisation and research [35].

The United Kingdom has taken a somewhat different approach, placing greater emphasis on developing the foundations that enable proportionate and adaptable regulation. Rather than introducing a single overarching AI law, the UK has invested in AI assurance, standards, testing capabilities and regulatory innovation. This includes initiatives such as the Regulatory Innovation Office [36], sector-focused regulatory sandboxes [37], on top of AISI, the AI Assurance Innovation and the Centre for AI Measurement. Together, these initiatives aim to improve the evidence base needed for future regulation while supporting innovation and deployment in parallel. The UK approach reflects a broader view that governance can be achieved not only through legislation but also through designated standards, codes of practice, technical guidance and market incentives. The evolution of the UK Government's Consumer IoT Security Code of Practice into ETSI EN 303 645 [38] provides a useful example of how policy objectives can mature into internationally recognised technical standards. More recently, ETSI EN 304 223 has established baseline cybersecurity requirements for AI systems, extending this model into the AI domain [39].

Similar approaches to the UK can also be seen internationally. Singapore has promoted assurance-oriented approaches through initiatives such as AI Verify [40], using testing frameworks and assurance tools to support trustworthy AI adoption rather than relying solely on prescriptive regulation. In the United States, voluntary frameworks, standards and evaluation capabilities including the NIST AI Risk Management Framework [4], the NIST AI Safety Institute Consortium [41] and work on AI TEVV [42], remain important parts of the assurance landscape. However,

federal policy has shifted towards a more explicitly innovation-led and light-touch approach: America's AI Action Plan prioritises innovation, infrastructure and international leadership [43], while Executive Order 14365 and the White House's March 2026 National AI Legislative Framework seek a nationally consistent regime that would pre-empt state requirements considered unduly burdensome. The legislative framework is not itself law, and states continue to develop requirements across areas including automated decision-making, discrimination, employment, healthcare and synthetic media, leaving organisations to navigate an evolving state-level compliance landscape.

Other countries have adopted different regulatory approaches that reflect their domestic policy priorities. Japan has taken an approach emphasising innovation, voluntary governance and international alignment. Its emerging framework combines non-binding guidelines, the AI Promotion Act [44] and sector-specific legislation, creating a flexible model that encourages organisations to adopt risk-based governance practices while avoiding highly prescriptive requirements [45]. This approach reflects Japan's longstanding preference for governance through guidance, standards and collaboration rather than extensive regulation. South Korea has pursued a middle path, introducing one of the world's first comprehensive AI laws through the AI Basic Act [46], while maintaining an explicit focus on innovation and international competitiveness. The framework adopts a risk-based approach with enhanced obligations for high-impact and generative AI systems, including transparency, safety and governance requirements, while also supporting AI development through national investment, standards and infrastructure initiatives.

China has adopted a distinct approach that combines industrial policy, extensive standardisation activity and targeted regulatory intervention [47]. As part of its objective to become a global leader in AI innovation by 2030, China has pursued a dual strategy of promoting AI adoption while introducing application-specific regulations for areas considered to present elevated societal or security risks. Rather than relying on a single comprehensive AI law, the regulatory framework has developed incrementally through targeted measures governing recommendation algorithms, synthetic media, and generative AI systems, alongside a rapidly expanding programme of AI standardisation [48]. This approach reflects a broader policy objective of balancing development and security, with the government acting simultaneously as a promoter of AI innovation, a regulator of high-risk applications and an active participant in national and international standards development.

An important feature of all these regulatory and non-regulatory approaches is the role of standards in translating high-level objectives into operational requirements. Such standards help define expected behaviours and controls, but generally leave open many of the measurement, testing and

evaluation questions needed to demonstrate conformance. They therefore create demand for assurance while reinforcing the need for robust measurement and evaluation methodologies.

Assurance methodologies

Between governance frameworks and technical evaluation activities sits an emerging layer of assurance methodologies, which provide structured approaches for assembling, evaluating and communicating evidence about the trustworthiness of AI systems. Unlike governance and risk management frameworks, which focus primarily on organisational processes and controls, assurance methodologies seek to answer a more specific question: what evidence is required to justify confidence that a system is safe, reliable and fit for purpose? These approaches therefore occupy an important intermediate position between high-level governance objectives and the practical generation of evidence through TEVV. Assurance methodologies operationalise assurance by providing systematic structures for gathering and interpreting evidence.

The **AI Security Institute (AISI)** occupies a distinctive position within the UK's emerging AI assurance ecosystem. Rather than developing governance frameworks or conformity assessment standards, AISI focuses on advancing the science of evaluating frontier AI systems. Through its evaluations programme, AISI has pioneered approaches for assessing the capabilities, limitations and risks of advanced AI models, including areas such as cyber security, autonomous behaviour, societal impacts, and misuse risks (a full list of publications can be found here [49]). AISI has helped establish the science of AI evaluation, contributing tools, methodologies and evidence that support future assurance, governance and standards development. AISI's work highlights both the importance and current immaturity of AI TEVV, reinforcing the need for rigorous approaches capable of generating reliable and reproducible evidence about AI system behaviour.

Many contemporary AI assurance methodologies draw on the concept of the safety case, which has been used for decades in safety-critical sectors including nuclear energy, rail, aerospace, defence and medical devices. A safety case is a structured argument, supported by evidence, that a system is acceptably safe for its intended use. Rather than relying solely on compliance with prescriptive requirements, safety cases require explicit justification of claims, the evidence supporting them, and the assumptions underpinning the argument. This approach recognises that complex systems cannot always be shown to be safe through testing alone and that confidence must be built through multiple complementary forms of evidence. The recent **BIG (Building, Interrogating and Governing) Arguments** approach for AI safety cases extends this tradition to contemporary AI systems [50]. Building on established safety-case methodologies, BIG provides a framework for constructing explicit assurance arguments around AI safety, identifying claims, supporting evidence, assumptions and uncertainties. The approach is particularly valuable for

advanced and generative AI systems where assurance depends on combining evidence from multiple sources, such as benchmarking, red-teaming, robustness testing and operational monitoring. Rather than treating assurance as a checklist exercise, the safety-case approach frames assurance as an evidence-based argument whose strength depends on the quality, relevance and completeness of the supporting evidence. A closely related development is the **Trustworthy and Ethical Assurance (TEA) framework** from the Alan Turing Institute and the University of York [51]. TEA extends the established safety-case and argument-based assurance tradition beyond safety to encompass broader trustworthiness properties such as fairness, explainability, accountability and contestability. Through an open-source platform, TEA supports multidisciplinary teams in constructing structured assurance cases that explicitly link claims about system behaviour to supporting evidence, assumptions and stakeholder requirements. By providing reusable argument patterns, evidence management tools and lifecycle-oriented guidance, TEA helps operationalise assurance as a continuous process rather than a one-off assessment. The **Assurance of Machine Learning for use in Autonomous Systems (AMLAS)** methodology is a specific implementation of this broader safety argument-based philosophy to the integration of machine learning within autonomous systems [52]. The methodology operates alongside the machine learning development lifecycle and provides a structured process for identifying safety requirements, assessing hazards, generating assurance evidence and constructing a safety case for machine learning components. A related methodology is **Safety Assurance of Autonomous Systems in Complex Environments (SACE)**, which extends safety-case thinking beyond machine learning components to autonomous systems more broadly [53]. Together, AMLAS and SACE represent some of the strongest examples of evidence-based assurance methodologies currently available, demonstrating how governance objectives can be translated into explicit claims and supporting evidence.

The Secure AI Framework (SAIF), developed by Google, represents a more specialised form of assurance focused on AI security [54]. SAIF was created in response to the recognition that AI systems introduce new classes of vulnerabilities that are not adequately addressed by conventional cybersecurity frameworks. These classes include threats such as data poisoning, model theft, prompt injection, adversarial manipulation and autonomous agent misuse. Rather than concentrating on broad trustworthiness objectives, SAIF focuses on identifying AI-specific security risks and the controls required to mitigate them throughout the AI lifecycle. The framework provides risk maps, security controls and assessment tools that organisations can use to understand and manage AI-related security risks. From an assurance perspective, SAIF demonstrates the growing emergence of domain-specific assurance frameworks that focus on particular attributes rather than attempting to address all aspects of trustworthy AI simultaneously. It also highlights how assurance increasingly requires evidence relating not only to model

performance but also to security, operational resilience and system interactions within broader socio-technical environments.

The Cross-Government AI Testing and Assurance Framework is a UK public-sector initiative developed by the Cross-Government Testing Community to provide a shared approach for testing, evaluating and assuring AI systems throughout their lifecycle [55]. The framework recognises that AI systems differ fundamentally from conventional software because they are probabilistic, data-dependent and subject to behavioural changes over time. The framework adopts a lifecycle perspective, covering activities from model planning and data preparation through validation, deployment and continuous monitoring, and highlights a range of trustworthiness attributes including fairness, explainability, transparency, robustness, performance, security and accountability. It promotes a modular testing strategy in which different evaluation activities are selected according to the characteristics, risks and intended use of the AI system rather than applying a single universal assessment process.

Singapore's AI Verify framework is one of the most developed attempts to operationalise AI assurance through a combination of governance principles, process checks and technical testing [40]. Unlike many governance frameworks, AI Verify explicitly links high-level principles to measurable outcomes and supporting evidence. The framework is organised around a set of internationally recognised AI governance principles, including transparency, explainability, robustness, fairness, accountability, security and repeatability. For each principle, AI Verify identifies desired outcomes, associated processes and evidence requirements that organisations can use to demonstrate responsible AI practices. A feature of AI Verify is its emphasis on transparency and verifiability. The framework combines documentary evidence with technical testing and reporting mechanisms, allowing organisations to provide objective evidence regarding system behaviour and governance practices. In doing so, it moves beyond purely process-based governance and begins to address some of the practical challenges associated with assurance. More recently, the AI Verify Foundation extended this work through the Global AI Assurance Pilot [56], an international initiative designed to explore emerging norms and best practices for the technical testing of real-world generative AI applications. Lessons from the programme are intended to inform future testing standards and best-practice guidance for AI assurance.

Collectively, these methodologies demonstrate how assurance arguments can be structured, but they also highlight the continuing need for consistent evidence-generation methods.

AI evaluation and assurance standards

The AI standards landscape is developing, with a large number of assurance related standards either published or currently under development across international, regional and national standards bodies. The AI Standards Hub (AISH) has become an important resource for coordinating engagement within the rapidly evolving landscape AI standards landscape. The AISH standards database references over 500 relevant AI-related standards in existence [57].

The standards described in this section differ from governance and management standards such as ISO/IEC 42001. Rather than defining organisational processes or management system requirements, they aim to provide the technical foundations, evaluation methodologies, testing processes and conformity assessment approaches needed to assess the properties and performance of AI systems. These AI specific standards are intended to be used within existing assurance methodologies, and in conjunction with existing mature domain-specific standards such as in biometric testing (ISO/IEC 19795), road vehicles (ISO 26262) or medical device software (IEC 62304). Much of this standardisation activity is coordinated through ISO/IEC JTC 1/SC 42 Artificial Intelligence, though other standards organisations active in this area include ITU, ETSI and IEEE.

The standards in this section can be grouped into four categories. **Foundational standards** define trustworthiness concepts and quality characteristics. **Technical measurement standards** provide metrics, measurement methods and evaluation methodologies for specific AI properties. **Process standards** define how testing, validation and assessment activities should be performed. **Conformity assessment standards** specify how evidence from testing and evaluation can be used to determine whether an AI system meets defined requirements.

Together, these standards form a progression from conceptual foundations, through measurement and testing, to assurance and certification of AI systems. Current work spans a wide range of technical topics including AI system quality, data quality, transparency, bias, robustness, testing, evaluation, impact assessment, conformity assessment and lifecycle processes. Table 2 provides an overview of some of the AI specific standards related to TEVV.

The ISO/IEC 42119 series is an emerging family of process standards focused on TEVV of AI systems, building on established software testing approaches, particularly the ISO/IEC/IEEE 29119 framework. The first published document, ISO/IEC TS 42119-2:2025, provides an overview of AI testing and adopts a risk-based approach to identifying appropriate test practices for AI systems. The wider series remains at an early stage of development, with additional parts under development covering topics such as verification and validation analysis (ISO/IEC DTS 42119-3),

red teaming (ISO/IEC TS 42119-7), and the assessment of generative AI systems (ISO/IEC TS 42119-8). As the series matures, it is expected to become a key component of the international TEVV landscape for AI assurance. At the European level, CEN-CENELEC JTC 21 is developing harmonised standards to support implementation of the EU AI Act, including standards for conformity assessment.

Despite the growing suite of AI evaluation and assurance standards, the landscape remains fragmented, reflecting both the breadth of AI applications and the relative immaturity of many technical assurance practices. While substantial progress has been made in defining governance requirements and risk management expectations, the TEVV ecosystem for AI remains comparatively less mature. Continuing gaps in AI evaluation methodologies, benchmarks and measurement science underscore the need for more robust evidence frameworks to support trustworthy AI [58].

Standard	Title	Description	Type	Status
ISO/IEC TR 24028	Overview of Trustworthiness in AI	Overview of AI trustworthiness concepts including robustness, reliability, safety, security, transparency, accountability and bias. Provides a conceptual foundation for AI assurance.	Foundational (TR)	Published (2020)
ISO/IEC 25059:2023	Quality Model for AI Systems	Defines AI quality characteristics such as functional correctness, robustness, transparency, intervenability and user controllability.	Foundational (IS)	Published (2023)
ISO/IEC TS 25058:2024	Guidance for Quality Evaluation of AI Systems	Guidance for evaluating AI systems against quality attributes defined in ISO/IEC 25059.	Process (TS)	Published (2024)
ISO/IEC 5259 Series	Data Quality for Analytics and Machine Learning	Defines data quality characteristics, measurement methods and governance requirements for AI datasets.	Technical Measurement Standard (IS)	Published series (2024-)

EN 18284	Quality and Governance of Datasets in AI	Defines requirements, metrics and methodologies for assessing data quality including representativeness, completeness and correctness.	Technical Measurement Standard (EN)	Draft
ISO/IEC TR 24029-1:2021	Robustness of Neural Networks Part 1: Overview	Overview of robustness concepts and existing methods for assessing the robustness of neural networks	Foundational (TR)	Published (2021)
ISO/IEC 24029-2:2023	Robustness of Neural Networks Part 2: Formal Methods	Methodology for assessing neural network robustness using formal verification methods.	Technical Measurement Standard (IS)	Published (2023)
ISO/IEC 24029-3	Robustness of Neural Networks Part 3: Statistical Methods	Methodology for assessing neural network robustness using statistical methods.	Technical Measurement Standard (IS)	Draft / Final approval stage
ISO/IEC TS 42119-2:2025	Testing of AI Part 2: Overview of Testing AI Systems	Applies software testing principles to AI systems using a risk-based testing framework.	Process Standard (TS)	Published (2025)
ISO/IEC DTS 42119-3	Verification and Validation Analysis of AI Systems	Provides approaches for AI verification and validation using simulation, formal methods and evaluation.	Process Standard (TS)	Draft
ISO/IEC TS 42119-7	Red Teaming	Guidance for AI red teaming, adversarial testing and robustness assessment.	Process Standard (TS)	Approved work item (pre-draft)
ISO/IEC TS 42119-8	Testing of AI	Definitions, concepts, requirements and guidance for assessing prompt-based generative AI systems.	Process Standard (TS)	Approved work item (pre-draft)

ISO/IEC 4213	Performance Measurement for Classification, Regression, Clustering and Recommendation	Standardised metrics for evaluating common machine learning tasks. Successor to the current classification-focused TS 4213.	Technical Measurement Standard (IS)	Draft
ISO/IEC 23282	Evaluation Methods for Accurate NLP Systems	Evaluation methodologies and metrics for natural language processing systems.	Technical Measurement Standard (IS)	Draft
EN 18281	Evaluation Methods for Accurate Computer Vision Systems	Standardised evaluation methods and metrics for computer vision systems.	Technical Measurement Standard (EN)	Draft
ISO/IEC 26582	Conformity Assessment for Functional Correctness of an AI System	Requirements and guidance for assessing whether an AI system performs its intended function correctly. Based on ISO/IEC 25059.	Conformity Assessment Standard (IS)	New Proposal
IEEE P8000.1	Assessment of the Trustworthiness of AI Systems	Methodology for assessing AI trustworthiness across accountability, robustness, transparency, fairness, privacy and oversight dimensions.	Conformity Assessment Standard (IS)	Draft

Table 2: An overview of AI evaluation and assurance standards. TR: technical report. IS: international standard. TS: Technical specification. EN: European standard. Status as of October 2026.

A recent development (2026) is the NIST AI TEVV Zero Draft initiative [42]. Rather than proposing a specific technical standard, the project seeks to establish a foundational framework that can underpin future standardisation efforts in AI testing and evaluation. NIST has explicitly recognised that current TEVV methods are highly diverse, vary considerably in maturity and continue to evolve rapidly. The purpose of the Zero Draft is therefore to develop a common conceptual structure, vocabulary and process model for TEVV activities that can accommodate different application domains and evaluation approaches while supporting future voluntary consensus standards. The initiative is intended to act as a bridge between existing evaluation practice and formal standards development, providing a common foundation upon which more detailed technical standards can subsequently be built.

The missing measurement layer

From the perspective of technical AI assurance, the regulatory layer can establish the legal requirements that organisations must satisfy and, through designated standards (or harmonised standards in the EU), it creates a pathway for translating those requirements into operational practices. In place of regulation, governments can also use non-statutory tools such as codes of practice or guidance to influence the market. For example, the UK government's 'Code of Practice for Consumer IoT Security' set core principles in Internet of Things security and the expectation on industry [38]. This code of practice then developed into an ETSI standard ETSI EN 303 645 [59]. The evolution of ETSI EN 303 645 from the UK Consumer IoT Code of Practice provides a useful example of how policy objectives can mature into internationally recognised technical standards. More recently, ETSI EN 304 223 has applied a similar approach to AI security, establishing baseline security requirements for AI systems while leaving open many of the measurement, testing and evaluation questions needed to demonstrate conformance in practice [39]. Assurance standards can therefore complement regulation or, where there is none, set expected industry practice, and drive market adoption regardless of regulation. Assurance frameworks provide an important mechanism for organising and communicating assurance evidence for assurance processes or other quality claims.

Nevertheless, a substantial gap remains between these frameworks and the measurement procedures needed to generate robust evidence. Even where standards define requirements for robustness, fairness or risk management, organisations must still determine how these properties should be measured, how uncertainty should be quantified and how results should be interpreted. As a result, tools like regulation and standardisation create the demand for assurance but have not yet resolved the challenge of producing reproducible, comparable and traceable evidence. This highlights the need for a robust measurement layer and evaluation methodologies capable of generating that evidence in a consistent and reproducible manner (Table 3). This was a motivation for the AI Verify Testing Framework and a similar conclusion led to the formation of the Appia Foundation, an international collaboration established under the Linux Foundation's Joint Development Foundation to develop practical mechanisms for demonstrating that AI systems satisfy regulatory, contractual and standards-based obligations [60]. Appia positions itself as the "connecting layer" between foundational standards and regulations on one side, and conformity assessment and assurance activities on the other.

Relatively little attention has been paid in assurance practice to these more fundamental questions of measurement, despite the fact that confidence in assurance outcomes ultimately depends on

the quality of the underlying measurements and evidence. As a result, many assurance activities rely on bespoke evaluation methods, specific benchmark datasets, locally defined metrics and context-specific interpretations of evidence, making it difficult to compare assurance outcomes or build confidence in results across different settings [61], [62]. Consequently, assurance remains highly dependent on expert judgement, limiting automation and scalability, comparability and confidence in results. This gap in understanding is in part because AI systems possess characteristics that differ fundamentally from traditional engineered systems. Many trustworthiness attributes, such as fairness, explainability, robustness or autonomy, are not directly observable physical properties but abstract constructs that must be assessed through proxy metrics [63]. Moreover, modern AI systems exhibit stochastic behaviour (which may be critical to system function, or something to minimise), dependence on data distributions, sensitivity to prompts and configurations, and performance variation across deployment environments. Foundation models and agentic AI systems introduce additional complexity through emergent behaviours and interactions that may not be predictable from component-level testing alone. The fact that many of their most impressive features are emergent properties, rather than being deliberately trained for and tested, means that assuring these properties and governing their use is critical to adoption and responsible development.

Principles / Values	OECD AI Principles, UNESCO Recommendation
Governance & Risk Management	NIST AI RMF, ISO 42001, ISO 23894, SAFE AI
Regulation & Compliance	EU AI Act
Assurance Methodologies	AMLAS, SACE, AI Verify, TEA
Technical Standards	prEN ISO/IEC DIS 23282, prEN 18281
Measurement Science (proposed)	Protocol libraries, traceability, uncertainty, comparability

Table 3: Layers of AI Frameworks. This table provides a layered view of example AI frameworks, assurance methodologies and technical measurement standards with examples at each layer.

Metrology is the discipline concerned with the theory and practice of measurement. It provides the principles, methods and infrastructure required to ensure that measurements are reliable, reproducible, comparable and fit for their intended purpose [64]. It is concerned with questions such as: *What exactly is being measured? How should it be measured? What are the sources of*

uncertainty? Can results be reproduced across organisations? Are measurements comparable across systems, domains and time? In traditional scientific and engineering domains, metrology underpins confidence in measurements by establishing clear definitions of what is being measured (the measurand), standardised measurement procedures, traceability to recognised references, and rigorous treatment of uncertainty. These principles allow measurements made by different operators, using different instruments and at different times and locations, to be meaningfully compared and interpreted.

Metrology for AI is the development of measurement principles for AI systems [65], [66]. Applied to AI assurance, a measurement science perspective shifts attention from simply reporting evaluation results to understanding the quality and validity of those results [67]. From this perspective, measurement is a structured process that generates evidence whose limitations, assumptions and uncertainties are explicitly understood. This evidence can be utilised in assurance frameworks and standards.

A fundamental role of measurement science is the development of a shared vocabulary and clearly defined measurands, ensuring that there is a common understanding of the property being assessed. A measurement science perspective therefore requires careful consideration of whether a chosen metric genuinely represents the property of interest and whether the resulting measurement is appropriate for the intended assurance objective. Measurement science also emphasises the importance of well-defined measurement and reporting procedures. Any evaluation result depends not only on the AI system itself but also on the data, test conditions, system configuration and operational environment used during assessment. Assurance therefore requires clear protocols specifying the datasets, model versions, prompts, hardware, software configurations and evaluation processes used to generate a result. Measurement science also provides for a systematic treatment of uncertainty including mechanisms for identifying sources of variability and understanding the confidence that can be placed in reported results. This provision of uncertainty helps ensure that apparent differences in performance are not simply artefacts of evaluation procedures, data selection or random variation. Traceability and reproducibility are also central metrological concepts. For assurance evidence to be credible, results should be traceable to the datasets, model versions, training procedures, prompts, random seeds, hardware and software environments used to generate them. Such traceability enables independent reproduction of results and supports meaningful comparison across systems, organisations and time. Without these capabilities, assurance findings are difficult to verify and may not be transferable beyond the original assessment.

Challenges for AI measurement

While we propose that metrology is an essential foundational layer for AI assurance, there are a number of challenges. Every AI system is unique. They are complex socio-technical systems highly dependent on their training data, model architecture, model training procedure and data workflow, deployment environment, user interactions and operational context. Two systems developed using the same underlying model may exhibit substantially different behaviours because of differences in inference parameters, privacy mechanisms, system integration, or operating environment. Even the same system may evolve over time through retraining, updates, changes in user behaviour or shifts in the operating environment. Consequently, assurance evidence generated for one AI system cannot automatically be assumed to apply to another, even where they appear superficially similar.

Fairness provides a particularly clear example. In some applications, assessing performance across demographic groups may be essential. For example, a healthcare system used for skin cancer diagnosis may need to demonstrate consistent performance across different skin tones, while an AI system used for recruitment may need to demonstrate that outcomes are not systematically biased against protected groups. However, the demographic dimensions that are relevant in one country or application may be entirely different in another. Categories considered important for fairness assessment in the UK may not reflect the demographics, legal requirements or social priorities of a different jurisdiction. Consequently, there is no universally applicable fairness metric; a recent study found over 60 fairness metrics in use in medical applications alone [68].

Staying within healthcare, AI systems frequently demonstrate strong performance during development but can perform differently when deployed in new clinical environments and may fail to impact healthcare outcomes [69]. For example, diagnostic models trained and evaluated using data from one hospital may experience reduced performance when deployed in another institution due to differences in patient populations, imaging equipment, clinical pathways, variations in environmental conditions or disease prevalence [70]. Consequently, a reported metric is often insufficient without understanding the evaluation population, uncertainty, operating conditions and intended use.

Financial AI systems present a different measurement challenge because the environment itself responds to the deployment of the model. Fraud detection systems, anti-money laundering tools and credit risk models operate in adversarial settings where criminal behaviour, market conditions and customer behaviour continually evolve. A model that performs well today may become

ineffective as fraud tactics change or as economic conditions shift. This changing background raises difficult questions regarding when a system should be re-evaluated, which populations should be monitored, and how performance degradation should be measured [71]. Similarly, a perception model used in an autonomous vehicle may achieve excellent performance under clear weather and daytime conditions but exhibit significantly different behaviour at night, in rain, fog or under sensor degradation [72]. Additionally, changes to system architecture, such as the addition of encryption, data fusion or safety mechanisms, can alter both system performance and vulnerability to adversarial attack.

This diversity means that AI assurance must be inherently context-dependent. Although some metrics can be considered task-specific, there is no single benchmark or threshold that can meaningfully characterise properties across all AI applications. The evidence required to assure a medical diagnosis system differs substantially from that required for a fraud-detection tool, an autonomous vehicle, a scientific assistant or a large language model embedded within a business workflow. Indeed, these requirements can also differ drastically between different tasks in the same domain. For example, a clinical diagnostic classification tool may be assessed against the performance requirements of a national screening programme. In this context, the key measurands may include sensitivity, specificity, and positive predictive value, reflecting the need to balance the risks of missed diagnoses against unnecessary referrals. This highlights that the choice of performance metrics depends on the intended use of the system. The assurance layer, therefore, must remain inherently flexible because system properties cannot be evaluated independently of context. This flexibility is critical for deployers looking to transfer a model from one problem to a similar situation, such as a disease classification model being applied to additional diseases. Despite the possibility of comparability, e.g. similar data types and deployment environments, issues (in addition to those previously mentioned) such as model fragility, adversarial robustness and ‘poisoning’ can become pertinent in these instances. Specific sector requirements also have a big role to play in AI assurance [73].

This context dependence challenges traditional metrology concepts applied to engineered products, whose performance can often be assessed against stable specifications and standardised test procedures. Measurement science for AI systems must therefore focus not on creating universal performance criteria, but on ensuring that evaluations are clearly scoped, the properties being assessed are explicitly defined, and the resulting evidence is valid for the intended application and operating conditions. The goal is not to eliminate contextual differences between AI systems, but to provide sufficiently rigorous and transparent measurement practices to the assurance layer so that results remain interpretable, reproducible and comparable.

Finally, National Metrology Institutes provide the highest level of reference for physical measurements, ensuring global comparability through common standards and international intercomparisons. This critical underpinning infrastructure ensures accurate and trustworthy measurements across the globe, from resilient timing signals for GPS and financial transactions, to precision manufacturing for aircraft construction, and the correct dose for medicines, and countless others. Not only does this traceability provide an unbroken chain to the highest point of reference, enabling technological advancements and resilience, but it also enables end users to determine the level of the traceability pyramid that is required for their application (e.g. a defined level of uncertainty). A natural question for AI metrology is what the equivalent reference should be for AI systems. While AI lacks stable physical artefacts against which measurements can be traced, analogous concepts may emerge through shared reference datasets, benchmark tasks, evaluation protocols and intercomparison exercises. Such mechanisms could provide the basis for internationally comparable AI assurance, allowing confidence in evaluation results to be established across organisations, sectors and national boundaries.

Protocols as a foundation layer for technical AI assurance

The Centre for AI Measurement (AIM) represents NPL's contribution to the UK's emerging AI assurance ecosystem, providing a focal point for research and development in AI measurement science. Its objective is to establish the measurement, evaluation and evidence-generation foundations required for trustworthy AI, building on the role that metrology has historically played in supporting quality, innovation and international trade in physical systems.

To address the challenges facing AI assurance and to provide a robust measurement layer, we propose the development of AI measurement protocols.

A protocol is a complete, repeatable and pre-declared procedure for generating technical evidence about one coherent property of a specified AI system under defined conditions.

A protocol describes the work step by step, from preparation and execution through analysis, qualification of the result and a final technical assertion.

A protocol should define:

- The property or question being evaluated and the measurand to be estimated
- The system under test, its configuration and the conditions to which the result will apply
- Required inputs, test materials, reference information and access
- The sequence of activities, decision points and permitted choices
- The analysis, metrics and treatment of variability and uncertainty
- Validity checks, limitations and conditions under which the result should not be generalised
- The evidence record and final technical assertion produced by the procedure

The purpose of a protocol is to generate interpretable and reproducible evidence. They are designed to provide practical, pre-normative technical procedures bridging frameworks and formal standards as they are still developing. Their intent is to provide characterised evidence for wider assurance cases and lifecycle decisions without replacing those frameworks. Protocols provide a practical mechanism for scaling AI assurance across industry, by allowing for a shared technical language between developers, evaluators, third-party assurers, regulators and users. This approach allows evaluators to state precisely which procedure, system configuration, data and conditions produced an evidence result. Protocols should be positioned alongside existing open

evaluation frameworks, such as AISI's Inspect framework [74], Stanford's HELM [75] or the AI Verify Testing Framework [40]. These frameworks encode evaluation components such as datasets, procedures and scoring methods in software, making assessments reproducible and reusable. A protocol overlaps with these functions but provides a broader measurement specification: it pre-declares the measurand, describes how measurement uncertainty will be treated and states the conditions under which the results are valid and applicable. Rather than replacing these frameworks, a protocol could reference these tools as the execution mechanism for part of the assessment.

Protocols also allow developers and vendors to make technical statements that can be independently verified within the stated scope and limitations. If widely adopted, they could support the comparison of results across systems, organisations and time, and reduce ad-hoc methodological choices by making deviations visible. Over time, widely adopted protocols can create the foundations for accreditation, certification and standards development, helping move the industry from fragmented, organisation-specific assessments towards a more mature and scalable assurance ecosystem. Perhaps most importantly, protocols balance consistency with flexibility: they provide a common measurement structure while still allowing evaluations to be tailored to specific applications, sectors, risk levels and regulatory contexts. This enables innovation to proceed without sacrificing comparability, transparency or trust.

It is important to note that protocols represent current recommendations, acknowledging that research in this area is ongoing. Protocols should therefore not be treated as static recipes; they may adapt over time and are evidence-generating instruments whose validity must itself be evaluated through case studies and assessment of utility to predict real-world system behaviour. Protocols themselves are not a conformity assessment. A protocol may apply an externally supplied threshold or decision rule, but the wider judgement that a system is acceptable, safe or compliant remains with the relevant assurance framework, regulator or decision owner, as this cannot be disentangled from the intended use.

The action of protocols in the AI value chain is shown in Figure 1 below. The Wardley diagram can be read from top to bottom to see how the customer-facing assured AI product depends on the underlying components, such as the AI system, its inputs, technical assurance, standards and datasets. The horizontal axis, read from left to right, shows how mature each component is, moving from novel or bespoke activities on the left to standardised, widely available capabilities on the right. The black bar represents inertia, meaning a barrier that prevents assurance practices from evolving from bespoke approaches towards more standardised, repeatable protocols. In this map, that inertia arises from fragmented requirements, such as differing standards, regulatory expectations, sector needs and interpretations of what constitutes sufficient assurance. The blue

MCR-005, Issue 1, July 2025

band represents the enabling role of protocols. It shows how protocols connect high-level standards to technical assurance, translating general requirements into specific, repeatable assessment methods that can generate evidence for an assurance case.

Protocols enabling a market for assured AI products

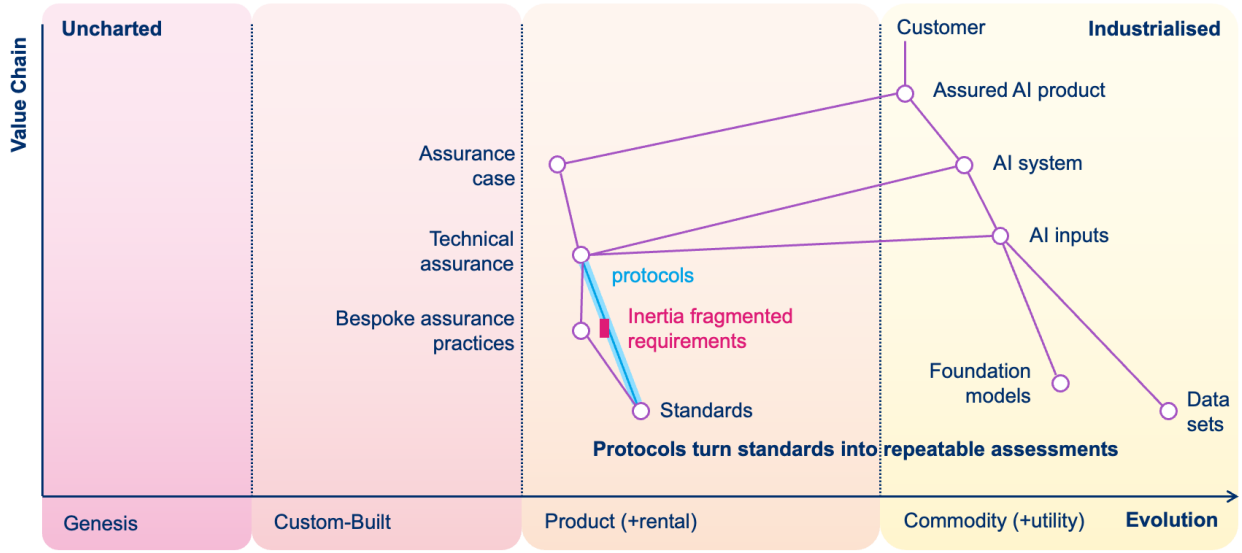


Figure 1: A Wardley diagram showing a simplified value chain for an AI product. The y axis represents the value chain of dependencies, while the x axis represents evolution of maturity. Only some of the inputs such as foundation models and data set are shown (NB. datasets could vary in their value). Protocols act to bridge the gap between standards and technical assurance.

A proposed protocol library for the measurement of AI systems

We propose that the development of a library of protocols will provide the range and depth of information needed to support assurance. The proposed protocol library is organised into a series of complementary layers that collectively support the generation, interpretation and maintenance of assurance evidence throughout the AI lifecycle (Figure 2). The structure reflects the distinction between measurement, evaluation, verification, validation and operational assurance, with each layer addressing a different aspect of evidence generation and confidence building. Rather than prescribing a single assurance process, the library provides a modular collection of protocols that can be selected and combined according to the assurance claim, system characteristics and deployment context.

The foundation of the library is formed by cross-cutting measurement methods. These protocols establish the measurement infrastructure required for all subsequent evaluations, including specification of the system under test, definition of measurands and evaluation designs, treatment of uncertainty and variability, technical traceability and reproducibility, qualification of evaluation datasets and graders, and the production of validity and applicability statements. These methods ensure that evaluation results are scientifically defensible, reproducible and interpretable, regardless of the property being assessed.

Building upon this foundation, the largest layer comprises pre-deployment technical evaluation protocols, which generate evidence regarding specific properties of AI systems. These protocols evaluate technical characteristics such as data quality, performance, robustness, safety, fairness, uncertainty, interpretability, privacy, security and system integration. The library also contains dedicated protocols for contemporary AI paradigms including generative AI, retrieval-augmented generation (RAG), tool-using agents and autonomous systems. Collectively, these protocols provide structured methods for assessing how a system behaves under defined conditions prior to operational deployment.

The protocol library then extends into operational monitoring methods, increasingly recognised as a key part of TEVV, making assurance a continuous activity rather than a one-off assessment [71]. These protocols support monitoring for data drift, model drift and concept drift, performance degradation, fairness changes, uncertainty shifts, behavioural changes, safety signals, operational incidents and system retirement. Their objective is to determine whether assurance evidence

remains valid over time and to identify conditions that should trigger re-evaluation or reassessment.

Two additional layers support the transition from evaluation results to assurance decisions. Verification protocols focus on confirming that systems have been built, configured and deployed as intended, including verification of software artefacts, pipelines, integrations and independent reproduction of technical claims. Validation protocols assess whether evidence generated during testing and evaluation is representative of the intended operational environment and whether the system is fit for its intended use. These protocols address prospective validation, human-AI workflows, local deployment contexts and controlled field trials. Together, verification and validation provide important links between technical evidence and real-world deployment decisions.

Finally, the library includes protocols for reporting and evidence publication, enabling technical evaluation results to be communicated consistently to procurers, regulators, assurance providers and other stakeholders. Supporting guidance and lifecycle assets complement the protocol library by providing protocol selection criteria, change-impact assessment approaches, vocabulary harmonisation and protocol-validation methodologies.

The protocol library will be maintained at NPL and will require a clear governance model defining the process of versioning and maintaining each protocol. This governance model will also establish transparent processes through which external parties can propose protocols, contribute evidence, challenge existing methods and request revisions, alongside an appropriate open licence that supports reuse while preserving attribution and version traceability.

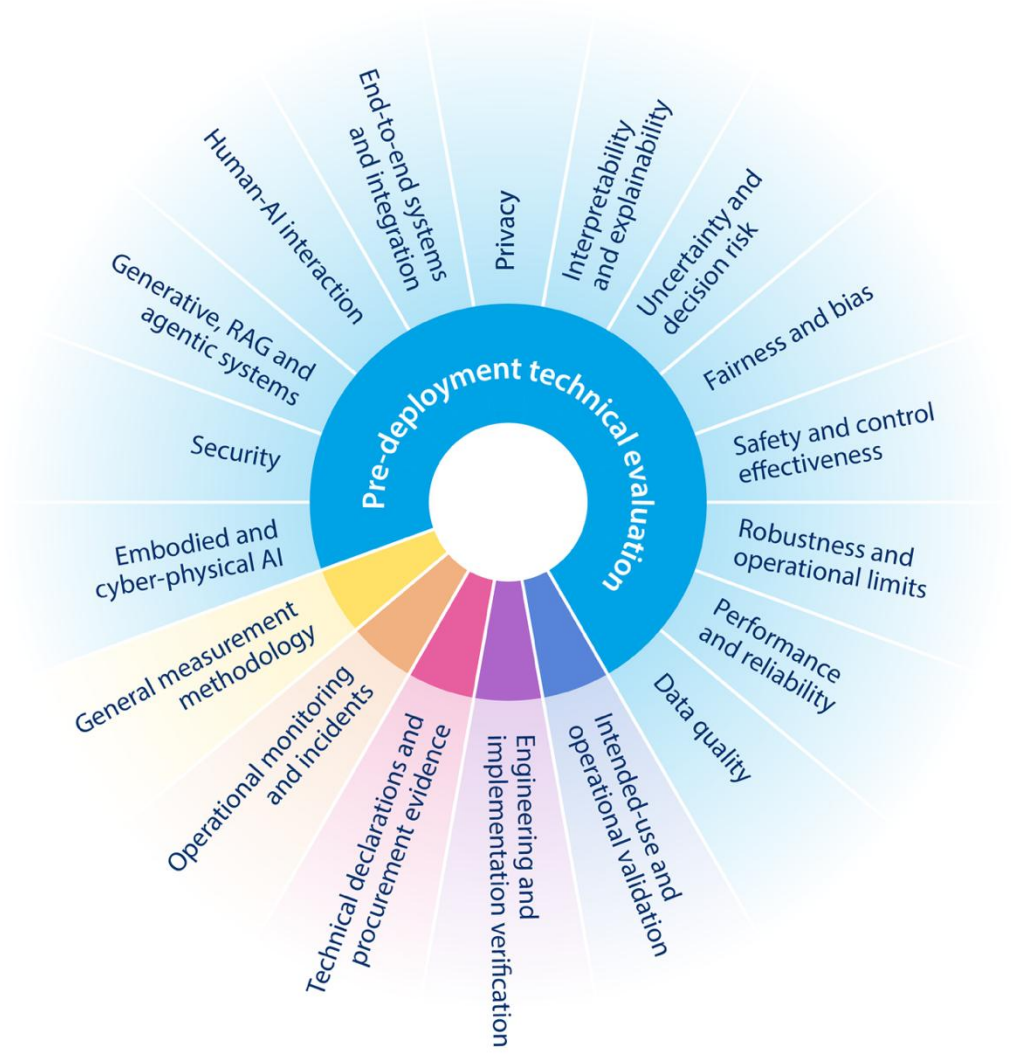


Figure 2: A proposed protocol library for the measurement of AI systems. Protocol areas spanning pre-deployment evaluation, validation and verification.

A worked example: a medical imaging assurance case

The TEA documentation from the Alan Turing Institute includes a case study for an AI-assisted diabetic-retinopathy screening system, designed to support practitioners reason over assurance questions and connect claims to evidence [76]. The example system analyses retinal images, produces clinical classifications, provides visual explanations and confidence information, and influences referral and monitoring pathways. The case study focuses particularly on whether explanations support clinicians in understanding and acting on recommendations. A broader assurance case for the same system could be decomposed into several property claims, each

supported by a different protocol package (Figure 3). Note that in a real case there would almost certainly be more claims and protocols required, including data quality and clinical assessment.

The example illustrates several important points. No single protocol demonstrates that the service is trustworthy. The goal claim is supported by a structured combination of technical arguments. Protocols within PR support the task-performance claim, while RO addresses robustness, FB addresses subgroup fairness, XI and UQ address explanations and uncertainty, and HI/SI protocols address the human and system workflow. The cross-cutting measurement protocols are invoked to make the property-specific evidence well specified, reproducible and appropriately qualified. Protocol evidence should be maintained as the system and context change and the conditions for reassessment and time-scale of the evidence should be specified through the general measurement protocols.

The example highlights the scope of the protocol approach. Applying a protocol does not establish that a goal claim is true. It supplies one characterised piece of evidence for an argument. It is the assurance framework that must determine whether the selected properties and evidence are complete enough for the intended decision and stakeholders. On the other hand, the same protocol evidence can serve several arguments, such as answering a regulator's question that no developer claim anticipated. Externally defined thresholds may be applied within a protocol result, but the justification for those thresholds and the final acceptability judgement remain with the assurance framework or decision authority. This separation of responsibility is necessary to allow for the myriad sector specific requirements on assurance cases. It is also important to remember that a technically correct protocol result may still be insufficient, irrelevant or outweighed by other evidence in the case.

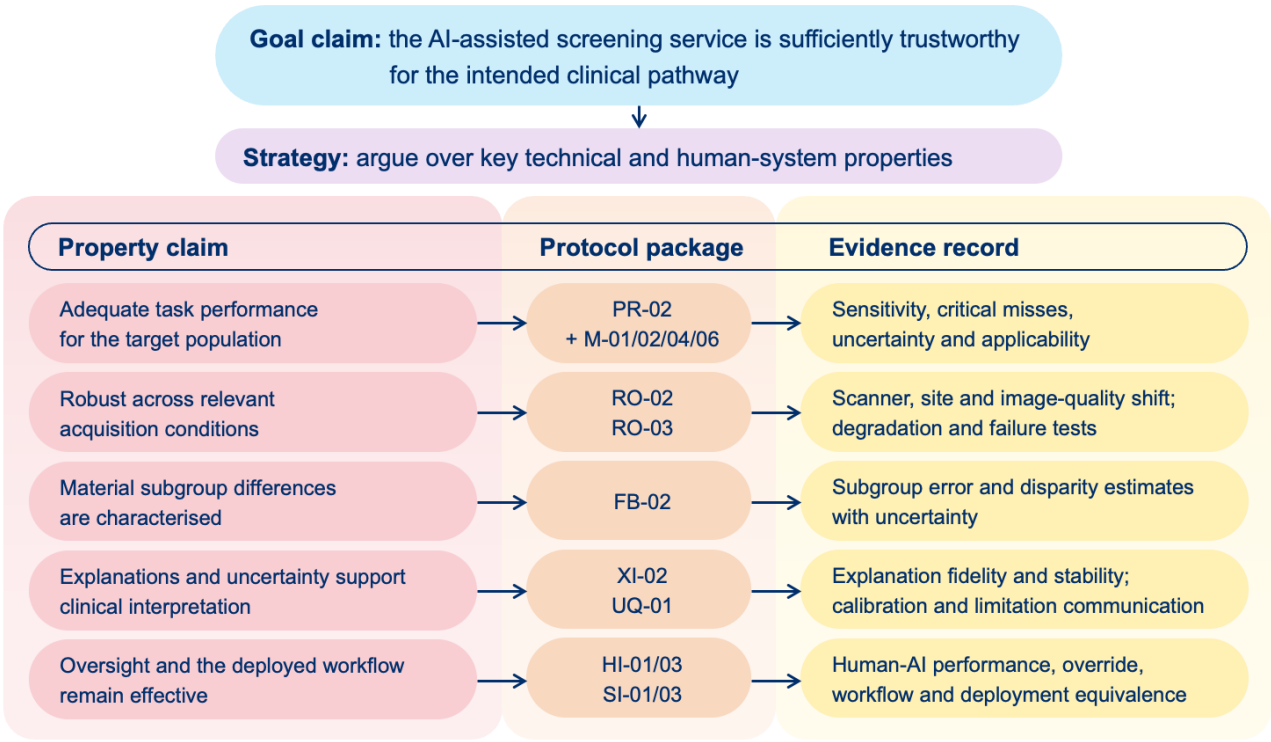


Figure 3: Illustrative decomposition of a medical-imaging assurance goal into property claims, protocols and evidence records. The example extends the ATI diabetic-retinopathy case study to demonstrate protocol selection. M: General measurement, PR: Performance, RO: Robustness and operational limits, FB: Fairness and bias, XI: Explainability and interpretability, UQ: Uncertainty quantification and decision risk, HI: Human-AI interaction, SI: End-to-end systems and integration.

From protocols to technical standards

A central premise of this report is that protocols can act as an intermediate layer between rapidly evolving research methods and formal technical standards. Standards development processes are necessarily conservative, consensus-driven and often require many years to mature. In contrast, AI technologies evolve on timescales of months rather than decades. Protocols provide a mechanism for capturing emerging good practice in a structured and repeatable form while retaining the flexibility to adapt as technologies, risks and evaluation methodologies develop. They therefore function as pre-standardisation artefacts, allowing technical communities to converge on common terminology, measurement procedures, reporting formats and evidential requirements before formal standardisation takes place.

Rather than emerging fully formed as standards, protocols can progress through a series of maturity levels. Early-stage protocols may be derived from academic literature, expert judgement and exploratory case studies. As they are applied in real-world settings they can be refined through pilot studies, cross-organisational evaluations and inter-laboratory comparisons. Over time, evidence accumulates regarding their repeatability, reproducibility, validity, discriminating power and operational utility. This creates a pathway from exploratory methods towards robust, widely accepted assurance practices capable of supporting accreditation and standardisation. A possible maturity model is shown in Figure 4. This progression mirrors the historical development of measurement methods in many scientific and engineering disciplines. New methods are typically proposed and refined through research studies, subsequently evaluated through case studies, collaborative exercises and inter-laboratory comparisons and eventually adopted by wider practitioner communities before being codified within national or international standards. Repeatability, reproducibility and fitness-for-purpose must be demonstrated before a method becomes suitable for conformity or standardisation.

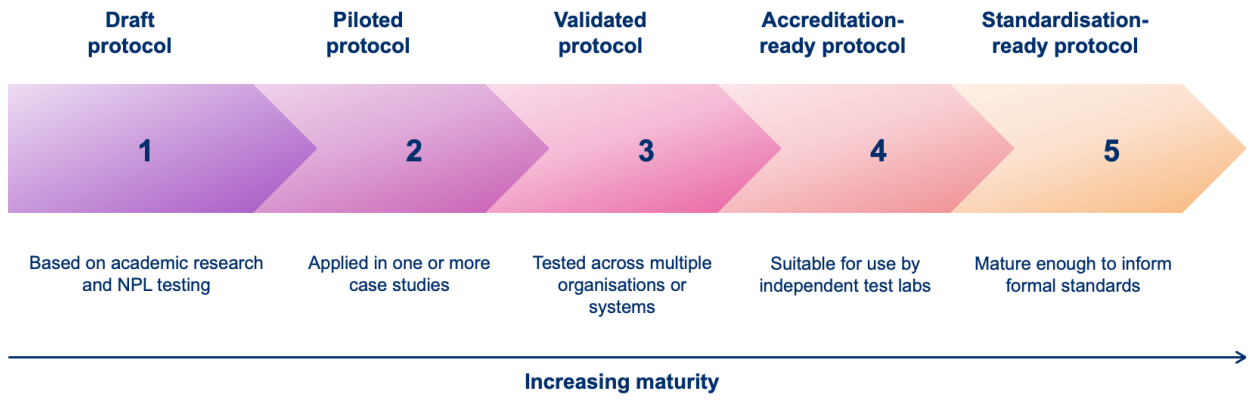


Figure 4: Protocol maturity levels leading from academic research to pre-standardisation.

Next steps for the Centre for AI Measurement

Looking ahead, the Centre for AI Measurement will focus on establishing the AI measurement protocols that provide repeatable, transparent and technically rigorous approaches for generating assurance evidence across different classes of AI systems. Where required, the Centre will build the scientific evidence base for AI measurement through applied research, intercomparison studies and peer-reviewed publications.

A major focus of the Centre will be to work with industry partners to identify and promote good practice for AI testing, evaluation, verification and validation (TEVV). These activities will include practical case studies and pilot implementations in priority sectors where assurance is particularly important, including healthcare, autonomous systems and financial services. These activities will help demonstrate how measurement science principles can be applied in real-world settings and provide evidence of the value of protocol-driven assurance approaches.

Findings from this work will be shared with the wider community and used to support the development of the third-party AI assurance ecosystem within the UK. The Centre will also contribute technical expertise to national and international standards development activities, helping to translate emerging measurement methods into future standards, conformity assessment approaches and assurance frameworks. In this way, the Centre aims to strengthen the foundations of the emerging AI quality infrastructure and provide confidence for the deployment of high-quality and trustworthy AI systems.

Summary

The future of AI adoption will depend not only on effective governance, regulation and standards, but on the ability to generate evidence that provides confidence in the quality, performance and reliability of AI systems. Protocols provide the practical bridge between high-level principles and measurable technical claims, while accreditation and standards provide the mechanisms needed to scale their adoption. Together they form the foundations of an emerging AI quality infrastructure, analogous to the systems of metrology, standards and conformity assessment that underpin quality, confidence and international trade across sectors. Establishing these foundations is therefore not simply an assurance challenge but a strategic opportunity to create the conditions under which organisations can deploy AI with confidence, markets can operate on a level playing field, innovation can flourish responsibly, and society can realise the benefits of AI through an ecosystem built on quality and internationally recognised evidence.

To realise this opportunity, coordinated action is required across the AI assurance ecosystem. Government should continue to support the development and adoption of a national AI quality infrastructure, including investment in measurement capabilities and creating incentives for the adoption of robust, evidence-based assurance. Standards bodies should work with measurement and assurance experts to translate high-level requirements into practical, measurable and reproducible standards. Assurance providers should adopt transparent, reproducible methods and contribute evidence on their validity and applicability. AI developers and deployers should engage early with assurance requirements and provide access to the systems, data and documentation needed for credible evaluation. NPL invites organisations across these communities to collaborate in piloting protocols, conducting inter-laboratory comparisons and building the shared evidence base needed to establish confidence in AI assurance.

Acknowledgements

We thank Louise Wright, Marina Romanchikova, Chris Burr and Adam Leon Smith for helpful discussions and for review of the manuscript.

References

[1] 'Introduction to AI assurance', GOV.UK. Accessed: Aug. 23, 2026. [Online]. Available: https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance

[2] 'Show us the evidence for the value of medical AI', *Nat. Med.*, vol. 32, no. 4, pp. 1163–1163, Apr. 2026, doi: 10.1038/s41591-026-04389-4.

[3] 'Artificial Intelligence Playbook for the UK Government (HTML)', GOV.UK. Accessed: Aug. 23, 2026. [Online]. Available: https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government/artificial-intelligence-playbook-for-the-uk-government-html

[4] I. T. L. Computer Security Division, 'NIST Risk Management Framework | CSRC | CSRC', CSRC | NIST. Accessed: Aug. 23, 2026. [Online]. Available: https://csrc.nist.gov/Projects/Risk-Management

[5] 'AI principles', OECD. Accessed: Aug. 23, 2026. [Online]. Available: https://www.oecd.org/en/topics/ai-principles.html

[6] *Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC (Text with EEA relevance)*. 2024. Accessed: Oct. 06, 2026. [Online]. Available: http://data.europa.eu/eli/dir/2024/2853/oj

[7] 'EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act'. Accessed: Aug. 23, 2026. [Online]. Available: https://artificialintelligenceact.eu/

[8] L. Floridi, M. Holweg, M. Taddeo, J. Amaya, J. Mökander, and Y. Wen, 'capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act', Mar. 23, 2022, *Social Science Research Network, Rochester, NY*: 4064091. doi: 10.2139/ssrn.4064091.

[9] 'ISO/IEC JTC 1/SC 42 - Artificial intelligence', ISO. Accessed: Aug. 23, 2026. [Online]. Available: https://www.iso.org/committee/6794475.html

[10] High Value Manufacturing Catapult, 'Trustworthy AI adoption Enabling safe and successful AI adoption in UK industry'. Spring 2026. [Online]. Available: https://hvm.catapult.org.uk/wp-content/uploads/2026/03/Trustworthy-AI-adoption.pdf

[11] J. Howell, 'Strengthening the AI Assurance Ecosystem', Partnership on AI. Accessed: Sep. 04, 2026. [Online]. Available: https://partnershiponai.org/resource/strengthening-the-ai-assurance-ecosystem/

[12] 'Assuring a Responsible Future for AI', GOV.UK. Accessed: Sep. 24, 2026. [Online]. Available: https://www.gov.uk/government/publications/assuring-a-responsible-future-for-ai/assuring-a-responsible-future-for-ai

[13] 'Trusted third-party AI assurance roadmap', GOV.UK. Accessed: Aug. 23, 2026. [Online]. Available: https://www.gov.uk/government/publications/trusted-third-party-ai-assurance-roadmap/trusted-third-party-ai-assurance-roadmap

[14] 'AI Opportunities Action Plan', GOV.UK. Accessed: Sep. 24, 2026. [Online]. Available: https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

[15] ‘Shaping Tomorrow: the UK’s Digital Standards Strategy (2026 to 2030)’, GOV.UK. Accessed: Sep. 24, 2026. [Online]. Available: https://www.gov.uk/government/publications/shaping-tomorrow-the-uks-digital-standards-strategy-2026-to-2030
[16] B. of Proof, ‘AI Risk Assessments Are Not Assurance — And the Difference Matters’, Burden of Proof. Accessed: Aug. 26, 2026. [Online]. Available: https://theburdenofproof.substack.com/p/ai-risk-assessments-are-not-assurance
[17] ‘ISO/IEC 27042:2015’, ISO. Accessed: Sep. 23, 2026. [Online]. Available: https://www.iso.org/standard/44406.html
[18] ‘ISO/IEC 27043:2015’, ISO. Accessed: Oct. 19, 2026. [Online]. Available: https://www.iso.org/standard/44407.html
[19] ‘The role of national metrology institutes in artificial intelligence and machine learning - NPL’. Accessed: Sep. 24, 2026. [Online]. Available: https://www.npl.co.uk/research/data-science-and-ai/national-metrology-institutes-in-artificial-intelligence-machine-learning
[20] ‘AI Safety Institute approach to evaluations’, GOV.UK. Accessed: Sep. 24, 2026. [Online]. Available: https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations
[21] ‘The AIQI Consortium’, The AIQI Consortium. Accessed: Aug. 25, 2026. [Online]. Available: https://www.aiqi.org
[22] ‘Ethics of Artificial Intelligence - AI | UNESCO’. Accessed: Sep. 23, 2026. [Online]. Available: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
[23] S. Ifayemi, E. Tabassi, and A. C. Deckard, ‘Decoding AI Governance: A Toolkit for Navigating Evolving Norms, Standards, and Rules’, Partnership on AI. Accessed: Sep. 04, 2026. [Online]. Available: https://partnershiponai.org/resource/decoding-ai-governance/
[24] ‘Playbook - AIRC’, NIST AI Resource Center. Accessed: Sep. 24, 2026. [Online]. Available: https://airc.nist.gov/airmf-resources/playbook/
[25] M. Levene *et al.*, ‘A Life Cycle for Trustworthy and Safe Artificial Intelligence Systems’. Accessed: Feb. 06, 2026. [Online]. Available: https://doi.org/10.47120/npl.MS57
[26] ‘Trustworthy and Ethical Assurance of Digital Healthcare’, The Alan Turing Institute. Accessed: Oct. 20, 2026. [Online]. Available: https://www.turing.ac.uk/research/research-projects/trustworthy-and-ethical-assurance-digital-healthcare
[27] ‘ISO/IEC 42001:2023’, ISO. Accessed: Aug. 24, 2026. [Online]. Available: https://www.iso.org/standard/42001
[28] ‘ISO 9001:2015’, ISO. Accessed: Aug. 24, 2026. [Online]. Available: https://www.iso.org/standard/62085.html
[29] ‘ISO/IEC 27001:2022’, ISO. Accessed: Aug. 24, 2026. [Online]. Available: https://www.iso.org/standard/27001
[30] ‘ISO/IEC 23894:2023’, ISO. Accessed: Aug. 24, 2026. [Online]. Available: https://www.iso.org/standard/77304.html
[31] ‘AI Management Essentials tool (accessible)’, GOV.UK. Accessed: Sep. 24, 2026. [Online]. Available: https://www.gov.uk/government/consultations/ai-management-essentials-tool/ai-management-essentials-tool-accessible
[32] ‘AI Risk Management Toolkit’, GOV.UK. Accessed: Sep. 24, 2026. [Online]. Available: https://www.gov.uk/government/publications/ai-risk-management-toolkit
[33] ‘Artificial Intelligence’, CEN-CENELEC. Accessed: Aug. 24, 2026. [Online]. Available: https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/

[34] ‘European AI Standardization | CEN-CENELEC JTC 21 |’. Accessed: Aug. 24, 2026. [Online]. Available: https://jtc21.eu/
[35] ‘Standardisation of the AI Act | Shaping Europe’s digital future’. Accessed: Aug. 24, 2026. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/ai-act-standardisation
[36] ‘Regulatory Innovation Office’, GOV.UK. Accessed: Sep. 25, 2026. [Online]. Available: https://www.gov.uk/government/organisations/regulatory-innovation-office
[37] ‘Regulatory sandboxes: what businesses can expect’, GOV.UK. Accessed: Sep. 25, 2026. [Online]. Available: https://www.gov.uk/government/publications/regulatory-sandboxes-what-businesses-can-expect/regulatory-sandboxes-what-businesses-can-expect
[38] ‘Code of Practice for consumer IoT security’, GOV.UK. Accessed: Sep. 24, 2026. [Online]. Available: https://www.gov.uk/government/publications/code-of-practice-for-consumer-iot-security/code-of-practice-for-consumer-iot-security
[39] T. Lawrence, ‘ETSI releases world-leading standard for securing AI’, ETSI. Accessed: Sep. 24, 2026. [Online]. Available: https://www.etsi.org/newsroom/press-releases/2627-etsi-releases-world-leading-standard-for-securing-ai/
[40] ‘What is AI Verify’, AI Verify Foundation. Accessed: Aug. 24, 2026. [Online]. Available: https://staging.aiverifyfoundation.sg/what-is-ai-verify/
[41] ‘NIST AI Consortium’, *NIST*, Jun. 2025, Accessed: Sep. 25, 2026. [Online]. Available: https://www.nist.gov/artificial-intelligence/nist-ai-consortium
[42] ‘NIST’s AI Standards “Zero Drafts” Pilot Project to Accelerate Standardization, Broaden Input’, *NIST*, Mar. 2025, Accessed: Aug. 25, 2026. [Online]. Available: https://www.nist.gov/artificial-intelligence/nists-ai-standards-zero-drafts-pilot-project-accelerate-standardization
[43] ‘White House Unveils America’s AI Action Plan’, The White House. Accessed: Oct. 21, 2026. [Online]. Available: https://www.whitehouse.gov/releases/2025/07/white-house-unveils-americas-ai-action-plan/
[44] ‘Act on the Promotion of Research and Development and Utilization of Artificial Intelligence-Related Technologies | e-Gov Law Search’. Accessed: Sep. 25, 2026. [Online]. Available: https://laws.e-gov.go.jp/law/507AC0000000053
[45] H. Habuka, ‘Japan’s Agile AI Governance in Action: Fostering a Global Nexus Through Pluralistic Interoperability’, Sep. 2025, Accessed: Sep. 25, 2026. [Online]. Available: https://www.csis.org/analysis/japans-agile-ai-governance-action-fostering-global-nexus-through-pluralistic
[46] ‘Korea AI Basic Act’. Accessed: Sep. 25, 2026. [Online]. Available: https://aibasicact.kr/
[47] ‘AI governance around the world – China’, The Alan Turing Institute. Accessed: Sep. 25, 2026. [Online]. Available: https://www.turing.ac.uk/news/publications/ai-governance-around-world-china
[48] ‘China Electronics Technology Standardization Institute’. Accessed: Sep. 25, 2026. [Online]. Available: https://www.cesi.cn/page/index.html
[49] ‘AISI Research & Publications | The AI Security Institute’, AI Security Institute. Accessed: Sep. 25, 2026. [Online]. Available: https://www.aisi.gov.uk/research
[50] I. Habli *et al.*, ‘The BIG Argument for AI Safety Cases’, Mar. 31, 2025, *arXiv*: arXiv:2503.11705. doi: 10.48550/arXiv.2503.11705.
[51] ‘Trustworthy and Ethical Assurance (TEA)’, The Alan Turing Institute. Accessed: Aug. 24, 2026. [Online]. Available: https://www.turing.ac.uk/research/research-projects/trustworthy-and-ethical-assurance-tea

[52] U. of York, ‘Assurance of Machine Learning for use in Autonomous Systems (AMLAS)’, University of York. Accessed: Aug. 24, 2026. [Online]. Available: https://www.york.ac.uk/assuring-autonomy/guidance/amlas/
[53] U. of York, ‘The Safety Assurance of autonomous systems in Complex Environments (SACE)’, University of York. Accessed: Aug. 24, 2026. [Online]. Available: https://www.york.ac.uk/assuring-autonomy/guidance/sace/
[54] ‘Secure AI Framework’, SAIF: Secure AI Framework. Accessed: Aug. 24, 2026. [Online]. Available: https://saif.google/secure-ai-framework
[55] ‘How to Test and Evaluate AI?’, How to Test and Evaluate AI? Accessed: Aug. 26, 2026. [Online]. Available: https://testing-ai-standards.github.io/cross-gov-ai-testing-framework/
[56] ‘Global AI Assurance Pilot’, AI Verify Foundation. Accessed: Aug. 24, 2026. [Online]. Available: https://staging.aiverifyfoundation.sg/ai-assurance-pilot/
[57] ‘AI Standards Hub | Advancing responsible AI through standards’. Accessed: Oct. 19, 2026. [Online]. Available: https://aistandardshub.org/
[58] ‘A five-step roadmap to closing the AI evaluation gap’. Accessed: Sep. 24, 2026. [Online]. Available: https://oecd.ai/en/wonk/a-five-step-roadmap-to-closing-the-ai-evaluation-gap
[59] ‘ETSI industry standard based on the Code of Practice’, GOV.UK. Accessed: Sep. 24, 2026. [Online]. Available: https://www.gov.uk/government/publications/etsi-industry-standard-based-on-the-code-of-practice
[60] ‘Appia Foundation’, Appia Foundation. Accessed: Aug. 26, 2026. [Online]. Available: https://appiafoundation.org/
[61] I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna, ‘AI and the Everything in the Whole Wide World Benchmark’, Nov. 26, 2021, *arXiv*: arXiv:2111.15366. doi: 10.48550/arXiv.2111.15366.
[62] H. Wallach *et al.*, ‘Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge’, Jun. 06, 2025, *arXiv*: arXiv:2502.00561. doi: 10.48550/arXiv.2502.00561.
[63] A. Z. Jacobs and H. Wallach, ‘Measurement and Fairness’, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, in FAccT ’21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 375–385. doi: 10.1145/3442188.3445901.
[64] ‘JCGM Publications: Guides in Metrology’, BIPM. Accessed: Jun. 24, 2026. [Online]. Available: https://www.bipm.org/en/committees/jc/jcgm/publications
[65] C. Welty, P. Paritosh, and L. Aroyo, ‘Metrology for AI: From Benchmarks to Instruments’, Nov. 05, 2019, *arXiv*: arXiv:1911.01875. doi: 10.48550/arXiv.1911.01875.
[66] T. Adel, S. Bilson, M. Levene, and A. Thompson, ‘Trustworthy Artificial Intelligence in the Context of Metrology’, Jun. 14, 2024, *arXiv*: arXiv:2406.10117. doi: 10.48550/arXiv.2406.10117.
[67] L. Wright, J. Clarke, V. Livina, C. Barnes, S. Thomas, and P. Duncan, ‘Getting the Measure of AI: Building Confidence in AI and Assurance through Metrics and Measurement. Bushy House Scientific Meeting, 21-22 October 2025’. Accessed: Aug. 25, 2026. [Online]. Available: https://doi.org/10.47120/npl.bhsmr1
[68] J. Matos *et al.*, ‘Critical appraisal of fairness metrics for artificial intelligence-based clinical prediction models: a scoping review’, *Lancet Digit. Health*, vol. 8, no. 7, Jul. 2026, doi: 10.1016/j.landig.2026.101001.
[69] ‘Show us the evidence for the value of medical AI’, *Nat. Med.*, vol. 32, no. 4, pp. 1163–1163, Apr. 2026, doi: 10.1038/s41591-026-04389-4.

[70] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, 'Key challenges for delivering clinical impact with artificial intelligence', *BMC Med.*, vol. 17, no. 1, p. 195, Oct. 2019, doi: 10.1186/s12916-019-1426-2.
[71] A. Rao *et al.*, 'Challenges to the monitoring of deployed AI systems: Center for AI Standards and Innovation', National Institute of Standards and Technology (U.S.), Gaithersburg, MD, NIST AI 800-4, Mar. 2026. doi: 10.6028/NIST.AI.800-4.
[72] N. Kalra and S. M. Paddock, 'Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?', *Transp. Res. Part Policy Pract.*, vol. 94, pp. 182–193, Dec. 2016, doi: 10.1016/j.tra.2016.09.010.
[73] 'Report: A Maturing AI Assurance Ecosystem: Sector Specific Applications'. Accessed: Aug. 26, 2026. [Online]. Available: https://www.techuk.org/resource/a-maturing-ai-assurance-ecosystem-mapping-sector-specific-applications.html
[74] 'Inspect', Inspect. Accessed: Oct. 20, 2026. [Online]. Available: https://inspect.aisi.org.uk/
[75] 'Holistic Evaluation of Language Models (HELM)'. Accessed: Oct. 20, 2026. [Online]. Available: https://crfm.stanford.edu/helm/
[76] 'Explainable Diabetic Retinopathy Screening System | TEA Documentation'. Accessed: Oct. 06, 2026. [Online]. Available: https://assuranceplatform.azurewebsites.net/docs/curriculum/case-studies/diabetic-retinopathy-screening

National Physical Laboratory

From Frameworks to Protocols:
A Measurement Science Foundation for Technical AI Assurance

Providing the measurement capability
that underpins the UK's prosperity
and quality of life

To find out more about NPL:
npl.co.uk

To get in touch:
npl.co.uk/contact
020 8977 3222

Hampton Road
Teddington
Middlesex
TW11 0LW

Please consider the environment
before printing this document

© NPL Management Ltd, 2026