

Bushy House Scientific Meetings Report 1

Getting the measure of AI: building confidence in AI and assurance through metrics and measurement

21-22 October 2025



National Physical Laboratory (NPL)

Getting the measure of AI: building confidence in AI and assurance through metrics and measurement

Louise Wright, Jennifer Clarke, Valerie Livina, Chris Barnes, Spencer 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.**

This report was prepared by NPL following the Bushy House Scientific Meeting (21-22 October 2025) and may not reflect the specific views of all individuals present.

© NPL Management Limited, 2026

<https://doi.org/10.47120/npl.bhsmr1>

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

Extracts from this report may be reproduced provided the source is acknowledged and the extract is not taken out of context.

Foreword

Artificial Intelligence (AI) has the potential to be transformative, with societal and economic implications felt worldwide. But it requires collective engagement to drive development and assure trust and confidence. In 1958, the National Physical Laboratory hosted one of the first international meetings on artificial intelligence, “Mechanisation of Thought Processes”, as part of its symposium series [1]. This meeting brought together multidisciplinary experts on topics of national or international importance where ground-breaking advances, more demanding requirements or a rapidly changing scientific landscape required in-depth discussion.

The 2025 Bushy House Scientific Meeting, “Getting the measure of AI: building confidence in AI and assurance through metrics and measurement” continued the spirit of those early symposia, bringing diverse perspectives together - government representatives, technologists, industrial leaders, academics, and experts from the quality infrastructure - to discuss confidence in AI. The discussions during the event were vibrant and wide-ranging, reflecting the complexity of today’s AI landscape including narrow machine learning models, generative AI, large language models and emerging agentic systems. This variety in approaches ran alongside the diversity of sectors and interests of those in attendance.

AI will not be shaped by one discipline alone, but by a rich interplay of ideas and expertise. This meeting, with its engaging agenda and space for discussion, fostered active engagement across disciplines. Throughout the meeting the participants shared a willingness to confront difficult questions about trust, assurance, and accountability. These conversations are essential because common language, with clear and transparent standards will be the bedrock of confidence in AI. Without clarity in terminology and transparency in practice, we cannot hope to build trusted systems that society will accept and rely upon.

Metrology - the science of measurement - may seem far removed from AI at first glance, yet it is fundamental to assurance. Testing, evaluation, validation, and verification all depend on robust measurement frameworks. The UK’s national quality infrastructure can provide a foundation for this work, and its role in AI assurance will only grow as technology

evolves. Establishing confidence in AI is not optional; it is critical to unlocking its benefits while managing its risks.

The insights captured in this report are not an endpoint but a starting point and provide a foundation for future consultation and action. They challenge us to think deeply about metrics, standards, and governance, and to work collaboratively across disciplines and borders. I hope this document inspires continued dialogue and the recommendations contained within can become concrete steps toward a trustworthy AI ecosystem, where AI can deliver for all.

Professor Dame Wendy Hall

University of Southampton

Executive summary

Artificial Intelligence (AI) is rapidly reshaping society and transforming expectations of the future. Yet as AI capabilities advance, the gap between frontier research and technologies that can be safely deployed in industrial settings has grown substantially. This gap is driven by the need for systems that are reliable, low-risk and manageable by non-specialist users, as well as by persistent uncertainty about how AI should be measured and assessed. Without clearer methods for evaluation, this gap risks constraining adoption of trustworthy AI, as well as AI's potential to support productivity and economic growth.

Ensuring the trustworthiness of AI requires coordinated action among developers, regulators, domain specialists and end users. While delivering the needs of end users can be addressed through rigorous, transparent and interpretable testing, regulators must additionally balance safety with practicality and resource constraints. These challenges span the entire AI system lifecycle; pre-market assurance and post-deployment monitoring often have different objectives and levels of complexity, and unclear funding responsibilities.

This Bushy House Scientific Meeting, Getting the measure of AI: building confidence in AI and assurance through metrics and measurement, focused on addressing these challenges through discussion with a broad range of stakeholders, including government, industry, academia and quality infrastructure. Within the discussion, six interconnected themes emerged:

- communication, education and training,
- human behaviour and socio-technical aspects,
- context, application and accountability,
- metrics for a system-level, outcome-focused approach,
- validation tools and techniques, and
- tensions between different stakeholder needs.

Together, these themes underscore that technical solutions alone cannot deliver trust; social, organisational and contextual considerations are also essential.

Communication and education were repeatedly identified as foundational. As a cross-cutting, multi-sector technology, AI requires shared understanding between technical experts, decision-makers and the public. Transparency is a prerequisite for trust, yet existing vocabularies, best practices and standards - such as those produced by the OECD, standards development organisations like ISO, and sector-specific bodies - remain inaccessible to many users. Effective adoption therefore demands consistent vocabulary, user-friendly tools, and training that supports safe operation, critical interpretation of outputs and informed judgement about when and how AI tools should be used.

AI systems also sit within broader socio-technical environments, shaping and being shaped by human behaviour. Their impact often extends beyond direct users, raising ethical considerations and questions of responsibility. Many stakeholders believe trust requires an ethical human in the loop, especially in sensitive or high-consequence domains. Explainability and transparency serve as partial proxies for this oversight - mirroring how we expect a person to justify their decisions - but they cannot fully replicate human reasoning or eliminate all sources of risk.

The meeting further highlighted the centrality of context. The same AI model may require different evaluation approaches depending on its application, risk profile and intended users. Trust thresholds vary significantly: safety-critical domains such as healthcare, security or defence demand far higher standards than lower-risk settings. Even when metrics are shared across sectors, acceptable thresholds differ based on the trade-off between correctness and the cost of errors. This reinforces the need for system-level frameworks that assess the full AI workflow - data, models, interfaces, human interactions and outcomes - rather than individual components. Regulation should therefore focus on application and intended purpose, while also considering system properties and user capability. Application-focused regulation can help clarify accountability in complex decision-making chains.

Current assurance practices draw heavily on academic research and existing metrics, but gaps remain between what these metrics measure and the realities of decision-making. Ethical considerations are particularly difficult to capture quantitatively. A more effective approach frames metrics as indicators of real-world outcomes, not just system outputs. Adopting a holistic systems view - incorporating algorithms, interfaces, users and external

context - enables more meaningful benchmarking and supports the development of metrics aligned with practical needs.

Consistency across the AI lifecycle is essential. Development often prioritises performance, while deployment focuses on risk-based criteria aligned with regulatory expectations. Regulators require traceability and transparency to support explanation and rectification, yet current regulatory guidance provides little clarity on appropriate tools or metrics. Without robust feedback mechanisms, practitioners may rely on inadequate methods and developers may not receive the insights needed to improve systems. Transparent governance on model suitability for regulated sectors can guide practitioner choices and help steer innovation toward trustworthy solutions.

Finally, the meeting recognised the tension between the rapid pace of AI innovation and the slower processes of standards development. Industry often resists independent testing due to cost concerns, while regulatory frameworks such as the EU AI Act increasingly favour risk-based, third-party assessment. Balancing innovation with safety remains a central challenge for the coming years.

Overall, the Bushy House Scientific Meeting highlighted the need for robust frameworks and methodologies that support assured, trustworthy AI. Priorities include developing standardised metrics for fairness, robustness and transparency; implementing continuous monitoring and validation pipelines; and establishing sector-specific standards supported by cross-industry collaboration. Future discussions should focus on data quality, privacy, infrastructure and the human-AI interface, with stakeholders encouraged to participate actively in shaping emerging standards and best practices.

Contents

1 Introduction and scope 9

2 Measurement for AI 11

3 What does trusted assurance of AI need?..... 18

4 Directions for the future: uniting the landscape for AI assurance30

5 References 33

6 Appendix - Acknowledgements..... 35

1 Introduction and scope

Artificial Intelligence (AI) has rapidly transformed society and the way we think about our future. In a short period of time, AI scope has accelerated, to include foundation models, open language models, large language models (LLMs) and generative AI. ChatGPT (developed by OpenAI) became the fastest growing app in history. We are also now living in a world where Agentic AI, which can act on its decisions via interaction with other software systems, is becoming more prominent. Although AI has a huge potential to benefit society, the development of AI poses many questions for the practice and theory of metrology (the science of measurement).

AI is already having a growing influence, both direct and indirect, on many aspects of daily life. Outside of the IT sector, the financial services industry has experienced the fastest growth and has been the biggest spender on AI services [2]. AI systems are used by businesses in the financial sector for trading and investment risk assessment, by clinicians for medical diagnosis, by industry for control of manufacturing processes and by the public for information-gathering. All these examples save time, money or both; but all also have the potential to cause financial losses or harm to people and the environment. To enable society to reap the benefits from this emerging and transformative technology, it is vital that they have confidence in its safe and trustworthy application.

When a product has the potential to cause harm to people, the government and society have a duty to ensure that the probability of harm being caused by the product (via its use or misuse), is understood and if possible, minimised. This understanding of the risk, and subsequent minimisation is usually communicated via some form of adherence to regulation, standards, certification or other form of assessment, for which a measurement approach is defined. The measurement approach is typically implemented as a set of physical tests and codified into standards that enable companies to provide appropriate evidence of their adherence. This approach, adopted across the world, has led to global approaches to standardisation (through bodies such as ISO) and mutual recognition of quality systems between different nations which eases trade and collaboration.

It is not currently obvious how to transfer this framework of product safety by certified test houses across to the AI domain. The current safety system assumes a static, physical

object that when tested will go to market and be used by consumers in the same form it was tested. Whilst software products are already subject to governance, the behaviour of an AI system depends on how its underpinning algorithms and machine learning models interact with the data it is exposed to. As a result, AI behaviour may diverge from expectations when used with data that vary from its training conditions, and the system itself may change over time if the algorithm is constantly re-training.

The range of AI models and algorithms is constantly changing, and the range of tasks that they can perform is constantly growing. In addition, the number of model parameters used within an AI system, the complexity of how the parameters lead to the outputs of the AI model and software, and the nature of the tasks AI is often designed to perform make specification of appropriate tests difficult. There are many metrics in use for assessment of AI models and algorithms that highlight different aspects of performance (for instance see the list here [3]) with acceptable performance depending on the specific use case.

Addressing the challenge of assessing the trustworthiness of AI requires a joint effort between a range of stakeholders, including developers, regulators, domain experts and end users. With that challenge in mind, the National Physical Laboratory (NPL) convened a Bushy House Scientific Meeting to discuss the challenge, starting from the following questions within two broad premises (“Metrics and benchmarks for trustworthy AI” and “Building Confidence in AI “):

- How do we balance AI adoption with a risk-based approach to AI development?
- What measurement standards do we need, and can we develop them?
- What types of metrics and benchmarks are needed in different sectors and for different use cases?
- How do we measure trustworthiness of AI and the quality of training data?
- Can we be confident in the interpretation of AI outputs by a non-expert user?
- What is needed to foster development, acceptance and uptake of robust AI assurance tools and metrics?

The meeting brought together 57 representatives from five countries, spanning government (21 %), the technology sector, industry (25 %), academia (12 %) and the quality infrastructure (30 %) for a wide-ranging discussion from AI evaluation and

assurance, to how to ensure the right metrics, quality measures and skills are in place to meet future challenges. A non-exhaustive list of attendees is available in the appendix.

This document is a summary of the discussions at the meetings. It sets out recommendations for effective AI assurance implementations and looks to the future for areas of focus.

2 Measurement for AI

This section of the document brings together the information and context from presenters that informed the discussions, it introduces specific challenges and frames them through the lens of measurement and assurance (The current state of play), and practical industrial deployment (In the wild: practical deployment).

2.1 The current state of play

The safe and secure use of AI systems is an international concern. Global summits have been convened involving the highest levels of government and key technology experts have been convened, such as the AI Safety Summit 2023 held at Bletchley Park in the UK. Governments across the world have developed guidance, policies, and in several cases, legislation aiming to support the use of AI whilst ensuring the risks to the populace are minimised. Many of these documents mention that systems should undergo appropriate levels of testing, typically dependent on the level of risk associated with the end use of the system. For instance, the EU AI Act [4] specifically mentions conformance testing and appointment of notified bodies (organisations approved to carry out conformance assessment of a given product or product class). However, in general little information is given to define what these tests might involve and what metrics are used to assess systems.

Quality infrastructure provides the backbone that underpins this risk management and ensures consumer safety; metrology (the science of measurement) provides the foundation for the development of tests, supports the codification of standards, and ultimately leads to tests carried out by accredited testing houses, enabling companies to provide appropriate evidence of their adherence. In response to the developing guidance, policies and legislation, and in response to the broader needs of industry, the global

quality infrastructure has begun to create standards, frameworks and guidance documents to define tests.

AI is a cross-cutting technology used in many sectors, and this has led to creation of standards that are sector-specific (vertical) and standards that are crosscutting (horizontal). A search of the Standards Database of the AI Standards Hub [5] in November 2025 returned 338 horizontal standards and 156 vertical standards, including items in development. This multiplicity of standards potentially leads to confusion for business since they may need to conform to multiple standards and the standards may not be consistent, particularly if the business operates across several sectors. The situation is further confused for companies operating in multiple markets since national standards are likely to vary as they may reflect the appetite for risk in that country. The cost of purchasing, integrating, and proving conformance with multiple standards may be a barrier for some companies, particularly small and medium enterprises.

Metrology is also rising to the challenge of AI. Metrology brings trust in measurement data through four main structures: stable globally agreed unit definitions, traceability chains, uncertainty evaluation and measurement validation by intercomparison. The agreed unit definitions are fixed and ensure that a kilogram (for instance) measured in the UK is the same as a kilogram measured in Brazil yesterday (or in ten years' time). This consistency ensures comparability of measurements made at different times and in different countries and underpins trade and scientific discovery.

Traceability chains (see figure 1) use a series of comparisons to ensure that the measuring devices used in the real world are consistent with the relevant unit definition. For instance, a National Metrology Institute (NMI) would hold the national standard for mass, traceable to the unit definition, and would use this standard in a mass comparator to calibrate masses belonging to calibration houses and other businesses, ensuring that those masses have a known value consistent with the national standard. Calibration houses would use their calibrated masses in mass comparators to calibrate the masses of their customers, and so on, and some of the calibrated masses would then be used to check the performance of balances in factories and in shops. This series of comparisons ensures that the measurements made throughout the chain can be linked back via the national standard to the unit definition.

In addition, NMI's also participate in international intercomparison exercises that check that claimed measurement capabilities in different countries are comparable. Uncertainty evaluation acknowledges that measurements are not perfect and are not perfectly repeatable: there may be a difference between the measurement result and the true value of the quantity of interest for reasons including thermal effects, operator skill, electrical noise, vibration, etc. Uncertainty evaluation uses statistical methods to produce a statement, often supplied as a confidence interval, of the likely range of the true value of the quantity.

Traceability (mass and derived units)

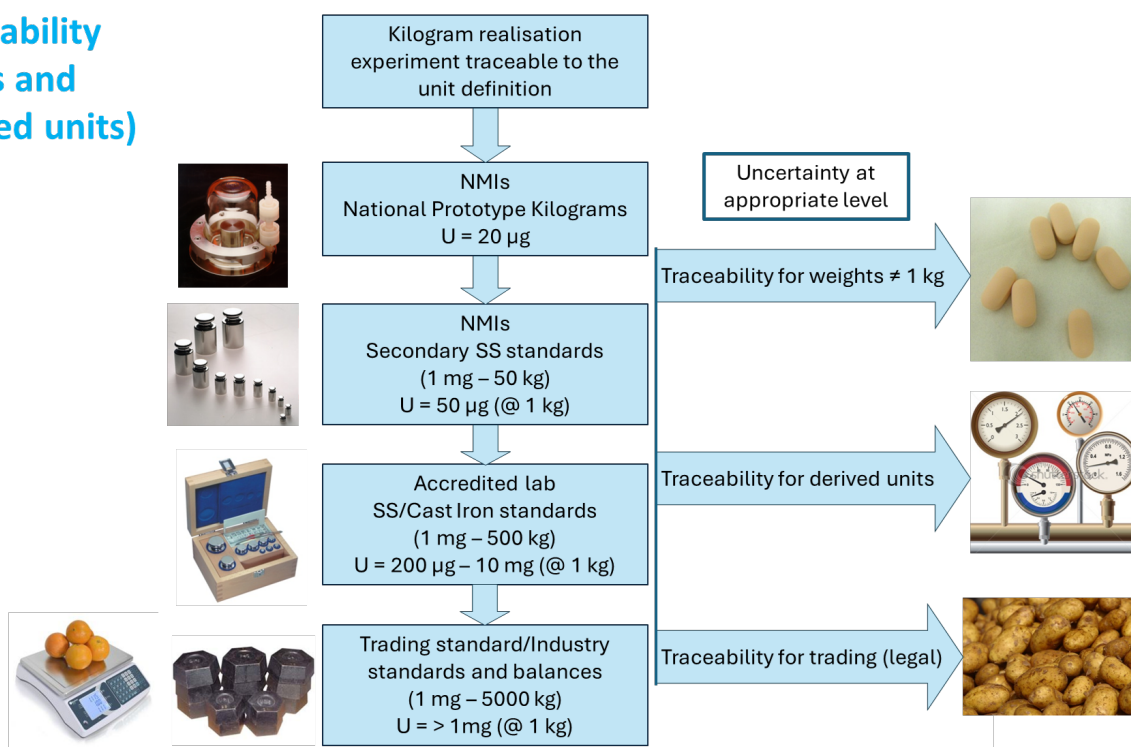


Figure 1: An example traceability chain for mass, showing mass ranges and example applications at each level. U here denotes the standard uncertainty associated with each stage in the chain.

Globally, NMIs are actively researching how to transfer this “traceability and uncertainty” approach to create trustworthy AI. As part of its activities on digitalisation and metrology, the Comité International des Poids et Mesures (CIPM, the global forum for NMIs that makes key decisions affecting metrology) has created a Task Group on Building Safe and Trustworthy AI. This group includes representatives from NMIs worldwide and is seeking to support metrology for AI, as well as the deployment of AI in metrology. A key aspect of its activities is input from stakeholders including the wider quality system, standards

organisations, and AI experts in general. There are also initiatives at regional level, for instance the EURAMET (the regional metrology organisation of Europe) network of mathematical and statistical experts has considered AI as part of their strategic research agenda [6]. As well as carrying out research on use of AI in metrology, some NMIs (such as LNE in France and NIST in the USA) are playing an active role as part of their country's AI safety institute (or similar organisation).

One of the central features that these quality system activities offer is justified trust, that is evidence-based trust provided by a neutral and independent organisation. This trust needs to be provided in unambiguous and well-defined language so that the nature of what has been assured is clear. For example, the EU AI Act refers to Safety and Security but in many European languages both these terms are represented by a single word leading to a lack of clarity. Terms such as risk management, governance and assurance are often used interchangeably even though they are distinct ideas, so a standardisation of concepts and vocabulary is needed. Work on this topic is already in progress, including work by NPL with OECD, but further development and dissemination is required.

Assurance and trust are necessarily context-dependent ideas. The level of assurance required for a safety-critical system will exceed that required for a lower risk system, but also each application will have specific performance requirements. For example, “false negatives” may be particularly problematic in a medical context, where they lead to a risk of disease going undiagnosed. Alternatively for example, “false positives” account for \$118 billion in retail losses and represent a large threat to the financial services industry [7]. A false-positive declined transaction works against the issuer and can result in erosion of customer loyalty. In addition to the specific performance requirements, the values and goals of the developer affect the way that algorithms are designed, and models are trained and tested, embedding contextual factors that strongly affect the AI system into its behaviour. In particular, data collection and consideration of end user needs may be easily affected by the values of the developer. Considering context within assurance processes potentially makes the use of generic approaches to assurance more challenging.

Finally, it should be noted that the end goal of many governments is to increase economic growth by deployment of AI at scale. A lack of trust is not the only barrier to deployment at scale. There is limited agreement on the skills required to use AI systems effectively and ethically in practice. The deployment of AI tools is a risk, and businesses may find that

placing a threshold on an acceptable level of risk is a challenge. Businesses may not have a comprehensive understanding of the risks associated with AI deployment in their sector, including the risks associated with unintended consequences, potentially adversely affecting long-term use of AI tools. Further, deployment at scale requires commitment at enterprise level, rather than at the level of individual users within that enterprise, which further complicates the definition of a risk threshold and leads to a gap between the decision-makers and those who see the direct impacts of a system, including benefits and harms.

2.2 In the wild: practical deployment

The role of government is to anticipate developments that will have future impact on the population and to mitigate risks if possible. In practice this often means enabling others to understand the risks and how to mitigate them (as with, for example cyber security). For AI systems, there is a range of stakeholders who will need to understand different parts of the risk profile via testing. Algorithm testing happens throughout the AI landscape of developers and users. Tech companies want to demonstrate the performance of their algorithms to gain market advantage. Organisations that deploy systems want to understand the strengths and weaknesses of the systems so they can make appropriate purchasing decisions. Academics and developers want to benchmark algorithms to assess innovative approaches. End users want to know that algorithms are fair, transparent and consistent. Assurance of algorithms is therefore not confined to regulatory contexts, and the frameworks, tools and resources used to support AI assurance will need to consider the priorities and understanding levels of a range of groups.

Assurance outside of the tech industry tends to follow publicly available best practice whenever possible, particularly in regulated industries that are accustomed to standardised workflows and provision of evidence. Uncertainty evaluation is not yet widespread amongst industrial users, but methods with uncertainty evaluation built in are under development and have been deployed in some areas of scientific research. Ideally, model governance is addressed through data sheets, model cards, and accountability cards at a minimum, with an aim for state-of-the-art evaluation using multiple metrics. However, with rapidly changing technology, the assurance approaches are also having to adapt. For instance, in some cases, awareness of tools such as model cards [8] may be present but guidance on how to prepare a good model card would be beneficial. At a more

fundamental level, there is a need for assurance tools that go beyond model cards to give insight into performance on downstream tasks and the trustworthiness of the training data and guardrails.

For many companies, trustworthiness means security (in its broadest sense) in deployment. The increasing cyber security risks associated with cloud deployment are driving some sectors away from software as a service and towards on-premises deployment of bespoke tools on the end user's own terms. This trend is likely to include specialist large language models (LLMs) as well as bespoke data analysis tools.

In many cases, AI development relies on single benchmark metrics because they are an easy way to contrast the performance of one algorithm with its competitors. In some cases (particularly in the UK) end users want to have a traffic-light system for visualisation of uncertainty and risk, although this leads to potential ambiguity regarding what “amber” signifies. At present there is little agreement on how to assess state-of-the-art when using multiple metrics that may give competing answers or rankings.

One further challenge associated with industrial adoption of AI regulation and guidance, is that they do not adapt well to complex systems. As we move towards the use of agentic systems (autonomous systems with proactive adaptation and control), which can carry out actions in response to prompts, and towards systems where multiple AI systems “debate” to find the best answer to a prompt, we find a verification gap. This gap is particularly noticeable where models are used for discovery as we cannot know ahead of time what the ground truth is, and in emerging fields of discovery consensus on even qualitative aspects of ground truth may not exist. Explainability approaches do not generally bridge this gap and the usefulness of explainability typically depends on the user's ability to understand the explanation given. A regulatory approach in many high-risk areas currently involves assessing the processes used to create and validate the product (including the documentation of those processes) rather than the product itself. For AI systems this approach would need to be adapted to address the potential weaknesses of product groups within a given sector, but a common framework for such assessments may be possible. This approach also makes adapting to new technologies straightforward.

Industrial deployment is where context is of key importance. An algorithm that will control robots on a factory floor may behave very differently in a sandbox environment compared

to a real environment. Specifically, the sandbox will replicate an idealised factory and may not reproduce elements such as floor roughness, temperature variability, changes over time, and random human behaviour. Furthermore, combinations of these individual elements that could lead to unforeseen consequences may not be represented, so for instance if a human trips on a rough floor, then they may cause a hazard for an autonomous robot at a time when the robot is unbalanced due to floor roughness. The assurance of such systems will need to reflect such differences. Sectors such as autonomous vehicles are seeking to manage these differences through specification of an “operational design domain” or ODD which defines the circumstances under which a system has been approved for use, which may be of benefit in other areas.

Regulation will need to adapt to the technology trajectory, in terms of both the evolution of algorithm capabilities but also the combination of technologies (e.g. robots, agentic systems, remote vehicles) that may occur in real-world applications. The gap between the frontier of AI and what is deployed industrially is generally very large, which means the regulatory aspiration is achievable. This gap exists because industry deployment usually needs technology that is sufficiently reliable, low risk, and can be managed and assessed by industrial users. The current uncertainties around AI and the lack of clarity on how to measure it force this gap to be much larger than in some other emerging technology areas, potentially limiting the ability of AI to drive industrial growth.

In general, end users want testing that is rigorous, transparent, interpretable and explainable [9], but regulators need to find a balance between safety and resource intensity. Development of testing and regulation is further complicated by the potential need to bridge pre-market and post-deployment regimes, where end goals and complexity may differ significantly, making funding sources and responsibilities unclear. In general, the question of who pays for the testing is still open: it should be borne in mind that systems such as the MOT certificate, which certifies conformance with environmental and roadworthiness standards for UK vehicles, required significant investment when they were first set up in order to reach the state today where the end user pays to get their system certified annually.

3 What does trusted assurance of AI need?

The scene-setting outlined above in Section 2 (Measurement for AI) led into a wide-ranging and open-ended set of discussions. Attendees drew on their experiences and their aspirations to highlight challenges and gaps in good practice, as well as highlighting existing work that could be built upon and disseminated more widely. The discussions were focused by providing initial thoughts and questions within two broad premises (“Metrics and benchmarks for trustworthy AI” and “Building Confidence in AI”).

The first premise included discussion of gaps and pitfalls as well as highlighting areas of current good practice, and sought to understand how fairness, transparency, reliability and security could be used to quantify trust. The second premise explored the concepts of confidence and trust and used case studies to explore how trust might be affected by end use and end users. It soon became clear that all the discussion topics that arose under these premises overlap and interact, with six key themes emerging:

- Communication, education and training;
- Human behaviour and socio-technical aspects;
- Context, application and accountability;
- Metrics for a system-level, outcome-focused approach;
- Validation tools and techniques ;
- Resolving tensions between different stakeholder needs;

The discussions within the two original premises have been merged and summarised below within these six key themes.

3.1 Communication, education and training

The fact that AI is a cross-cutting technology with applications in multiple sectors inevitably means that appropriate communication is key to successfully generating trust.

“Appropriate communication” here encompasses not only a suitable vocabulary but also a suitable volume of information presented in an appropriate format. For instance, a neural network AI model supporting analysis of X-rays for diagnosis that is explainable should

show the radiographer the analysed image with highlighting or circling to indicate the parts of the image that most influenced its output, but should not present the user with unnecessary or unhelpful information such as the structure of the AI model and its weights and parameter values. The variety of potential end users of AI, from the wider public to tech researchers, means that this suitability will need to consider the end user as well as the application. There is a need for a body to act as a place to pool the learnings from different sectors and share the relevant information more broadly with industry as a whole, using appropriate language throughout.

An area for future development is a well-defined vocabulary. Some terms, such as “confidence” and “trust”, are formed from multiple dimensions with different aspects being important in different areas. A taxonomy of such terms will help to clarify what the needs of a given application are. Similarly, terms such as “reproducibility”, where both “explainability” and “interpretability” are important, require further clarification. Even the terms used to describe AI systems require clarification. For instance, “algorithm”, “system”, and “model” are often used interchangeably but there are differences between them that have significant implications for measurement and regulation. Whilst vocabularies already exist (for instance the OECD has a glossary of terms and ISO have issued a vocabulary standard and many sector-specific groups are working on similar resources), they are not readily understandable to all audiences.

Some aspects of AI systems still lack vocabulary definitions that would enable risk to be more adequately described and accurately assessed. One important example is the level of human involvement in each decision, where the ability to draw distinctions between agentic systems and autonomous systems, triage systems, and systems providing information to a human decision-maker would enable appropriate tests to be defined for each group.

Stakeholders with technical knowledge, and particularly regulators, face a combination of challenges. They commonly know that metrics measure performance and that different metrics measure different things, but they lack the knowledge to a) be confident that they are interpreting metric values correctly, and b) be able to select the right combination of metrics for a given problem. In some cases, they may also lack agreement on what aspects or understandings of “performance” are important to measure. Often the link between the measurement and the outcome is missing. A related challenge is “how good

is good enough?": what values of a metric should be deemed acceptable? It is also not clear how widely a given value assignment can be used: how similar do two applications need to be for the same metric values to be appropriate? Similarly industrial companies need guidance on selection of a suitable standard, since many cross-cutting and sector-specific standards exist. Some forms of algorithm, particularly generative AI systems that take in verbal prompts, may respond differently to different inputs and so prompt engineering skills will be needed in some sectors. All these aspects require well-designed training to address the key questions in a sector.

Finally, the volume, content, and presentation of communication must be chosen well. It is possible to give the illusion of transparency by swamping people with unusable and extraneous information: issuing the weights of a large-scale model does not make its decisions any more transparent. There is a need for well-designed communication of uncertainty, and particularly what it means for performance guarantees. Ultimately, if the AI is not transparent, it may not be considered trustworthy.

All of the topics touched on in this section, and the sections below, cannot be delivered solely as documents, requirements and guidance. People will need training and education to transform their behaviour and to embed an understanding of how to use AI tools effectively and safely, and how to interpret and critically assess results. This skill development will need to sit alongside the development of appropriate communication methodologies with commonly agreed terminology.

3.2 Tensions between different stakeholder needs

Following on from the need for a common vocabulary, the definition of "confidence" may differ for different users. Some users may want a measure of trust that they can assign to the result ahead of decision making (e.g. 60 % chance of containing an anomaly, prompting further assessments), whereas others may want to blindly follow the outcome (e.g. low risk high volume tasks). Further tensions arise from different aspects of trustworthiness, for instance use of synthesised data can lead to privacy problems when used to address unfair data. The use of synthetic data also prompts several other questions around potential sources of uncertainty, e.g. is the synthetic data generated by a trustworthy model? Has the model been trained on biased data, relative to the (test) population of its intended use?

AI models are contingent on the data they are trained on, and how data is collected can introduce bias due to privacy concerns, poor acquisition standardisation, combining different data sources (which may not be independent of each other), etc. In relation to privacy, GDPR seems to be poorly observed in some areas and to not have an enforcement mechanism. In general privacy and transparency are viewed as being in conflict and a trade-off is required, the “optimal” point varying with use case.

At present a lot of judgment is based on the answer to “is an AI better than a human”, which is a multi-dimensional problem, with “better” performance measurable in several different ways. Many classification cases examine metrics such as accuracy, sensitivity and specificity to compare AI to human experts, but throughput (i.e. the number of items passing through a process or system) may be an important metric to consider. Evaluations of AI vs humans need to reflect the intended use case in terms of data availability, quality and variability. If an AI model is 99 % accurate compared to 85 % for humans but cannot generalise to other test cases this is an unfair comparison. Moreover, in some cases “is an AI better than nothing” may be a more important question. Are we prepared to accept an algorithm that is 2 % worse at identifying faulty parts than a person if it processes them 100 times quicker so we can make more? The role of edge cases or out of distribution data may not be well tested or flagged by the models, and may need to be considered as part of any validation regimen, along with the consideration of an ethical human to validate outputs.

Balancing the need for standards for safe innovation (typically a slow process) to keep pace with rapidly developing technology causes friction between industry and standardisation bodies. For example, developers and testers may be lagging behind hackers leading to cyber security concerns. There is lobbying from industry against third party testing due to implication of cost and potential increase in time on them, whereas regulation (e.g. the EU AI act) takes a risk-based approach which may suggest independent testing.

3.3 Human behaviour and socio-technical aspects

AI systems do not exist in isolation: they interact with, and impact on, people. The impact frequently spreads beyond the group of end users. For instance, the end user of an algorithm for assessing X-ray images is a radiologist, but their patients will be strongly

affected by the algorithm's performance. The introduction of AI systems requires not only an assessment of the likely impact on the full range of stakeholders, but also careful design of the system and its interface to try to ensure that the impact is positive, fair and unbiased (or in some cases appropriately biased) for all of society. It is also important to embed testing of the algorithm and the interface by end users into the design process to reduce the probability that unexpected effects do not arise during deployment. The potential impact may go well beyond what system designers anticipate: it has been demonstrated that an algorithm can determine the patient's race from X-rays [10] where a human may not consider this aspect or have this knowledge.

Looking beyond bias, it is important that the end users are not adversely affected by their use of AI algorithms. One common concern is the possibility that people may fail to acquire, or even lose, knowledge and become "detrained" as they outsource tasks to AI. Another potential problem arises when systems are used as triage tools, for instance an algorithm to triage medical images might divide patient scans into "definitely ill", "definitely not ill" and "not clear", with the "not clear" category being examined by a human to reduce risk to patients. This approach results in the human only seeing cases that are difficult to assess, which can lead to cognitive burnout and is likely to make giving new staff on the job training more difficult.

It is widely understood that biased or unbalanced training data leads to a biased model. Data gathering is subject to cultural and social biases in addition to the bias that arises from limited metadata and data availability. Privacy concerns may further reinforce this availability bias. The question of use of copyrighted materials also remains unresolved. In addition, there are situations where bias in data is desirable. The prevalence of diseases such as sickle cell anaemia are affected by race, so a trustworthy algorithm may require a racially biased training data set.

One challenge with consideration of these socio-technical aspects is that there is no such thing as a standard human. The way that people interact with AI will potentially depend on factors such as cultural influence, age and educational level, but these dependencies are not likely to be clear-cut or easy to quantify. In addition, people may interact with systems differently on different days, both in terms of how they ask questions of the system but also in terms of how they interpret its outputs.

For many people, trust requires an ethical human in the loop. Acceptance and understanding of AI products are affected by the language used to describe them, so for instance “self-driving cars” are likely to be thought of differently to “cars with high levels of AI driver assistance”. Trust is further enhanced by end users developing experience with a tool and understanding its limitations. To some extent, factors such as explainability and transparency are attempts to address the lack of human in the loop: we could ask a person to explain how they reached a decision and could critique their reasoning, so we should ask the same of an algorithm. Transparency around training data can illustrate how the algorithm is interacting with wider society. However, these approaches have limitations. Explainability may have persuasive capability that could become manipulative and may not always enhance human performance.

It is important to consider these aspects in testing development. There are opportunities for developing underpinning frameworks such as standardised vocabulary for, and a taxonomy of, the ways in which people interact with systems. The rise of agentic systems is likely to lead to people being further removed from the steps that led to the information they are presented with, and being able to ensure that this type of system is tested appropriately is key to safe deployment and trustworthy AI.

3.4 Context, application and accountability

A recurring theme that emerged during the meeting is the importance of context in the assessment of AI, with the same AI model used in a different context potentially requiring distinct evaluation criteria. Fundamental considerations vary with context, such as how the data map to the real-world scenario and what is confidence? These aspects may also differ within the same application at different stages of the AI life cycle, for instance during development, deployment and procurement. Context will also impact the required levels of trustworthiness, which will be much higher in safety, security and defence applications compared to low-risk applications.

Although many commonly used metrics exist, they need to be output-specific and task-specific to evaluate practical models. Most industries actively want to use the right metric as there is a benefit to them (market advantage if they are external developers, lower risks if they are in-house teams). Metrics need to reflect the real-world context or scenario, and evaluation should be done in an operational setting (or appropriate proxy). For instance,

evaluation of autonomous systems in a lab setting is a useful preliminary step, but ultimately of limited value for deployment due to the high-risk application and dynamic nature of the surrounding environment. The operational context may include the interaction between the human and the machine because people are part of the system, so this consideration goes beyond metrics.

Another contextual challenge is “how good is good enough”? It is difficult to scale or transfer this information on metric levels across a sector, let alone between sectors. Even if the same metrics are used (either within a sector or in different sectors), the acceptable threshold may differ, ultimately depending on trade-off between the value of getting it right and the cost of a mistake. The risk mindset varies from sector to sector and may change according to the nature of the risk. For instance, people may have a zero-tolerance approach to cybersecurity risks due to the high cost of a mistake, whereas in other areas, the risks of undesired outcomes may be more tolerable. Further still, different users in the same context may have different risk appetites, e.g. the financial risk appetite of users may depend on their financial situation, thus the different notions of “good” should be based on what the AI system is intended to be used for and by whom. The application and outcome-specific nature prohibit the universal standardisation of metrics. Abstracting this to a framework level that considers the entire workflow, not just the data or model, can support cross purpose translation and is necessary for an assessment of fairness.

A lot of discussion centres around regulating AI and the challenges with future proofing this for a rapidly developing technology and defining exactly what is and is not AI. Instead, regulation should be focused on application and not on technology, but this approach needs to consider the properties of a system (e.g. deterministic or stochastic), intended use, and end users. Application focused regulation can address ambiguities in accountability when using AI. When used as a decision support tool (human in the loop) accountability can potentially remain with the human user making the decision, when used in an automated system (human out of the loop) then accountability may lie with the AI vendor and organisation using that system (e.g. ensuring regular Quality Assurance/Quality Control).

3.5 Metrics for a system-level, outcome-focused approach

A fully deployed AI system includes not only the algorithm but also the user interface for inputting data or questions and the interface that presents results to the user, and potentially any interfaces through which the algorithm controls other systems (e.g. an autonomous car). Assurance needs to include assessment of all these aspects. The status quo of assurance is built on academic research and development, the current literature base, and best practice around evaluation and metrics. Metrics are typically focused on a single aspect of the performance of a single AI algorithm and are often generated by the developers who know the system well. However, there is often a gap between what an accepted metric captures and the needs of decision-making in each area (for instance, can ethical factors even be measured meaningfully?).

Some metrics are needed to prove that systems cannot do specific things, so for instance demonstrating the training data does not contain unsuitable images proves that your algorithm does not cause harm caused by exposure to unsuitable images. Assessing in this way allows the hypothesis to be falsifiable, i.e. you only need to observe a single instance of an unsuitable image to demonstrate the potential to cause harm. This approach to assessment is useful for testing systems against rare events, counterfactuals, or fairness, and the impact on population and sub-groups. The development of these tests typically involves asking “what would you expect” and “what are you seeing” for a given system and identifying or developing a quantitative measure of the comparison between the answers to those two questions, benchmarking performance against expectations.

A further challenge of metric definition will also relate back to the wider context of AI system use, and how the problem and metrics are framed. The metrics ideally need to start from the stakeholders, with a clear distinction between benchmarking for performance and benchmarking for the regulatory environment. Benchmarking for performance is typically carried out by developers and may focus on a small group of metrics and the impact on direct end users, where the regulatory environment will require a wider set of metrics and will look at the wider group of people affected by the output of a system. Considering metrics as the end outcome of the use of the system, rather than outputs of systems, is a productive place to intervene.

There is an increasing need to differentiate assurance requirements between groups of AI systems. For example, within AI-augmented systems, the AI aspect may build on trusted mathematics or tested software, therefore requiring a different approach to assessment when compared to a purely AI system. The integration of multiple AI systems, reference data sources or agents, can lead to very different evaluation challenges, and metrics, compared to those for single AI systems. A particular challenge is uncertainty quantification within these integrated systems, and this is where an understanding of AI system sensitivity, error identification, and error correction will have a larger role. Additionally, there is a growing tendency to use LLMs to assess LLMs, and to focus on bias in input data when outcomes are arguably more important. AI systems, and reference data are also unlikely to be static, either through continuous or scheduled updates, and so metric values will change over time. Therefore, being metrics-driven rather than outcome-dependent can lead to “gaming the system”, as well as growth in energy/resource consumption for marginal gains.

If the general public are to trust AI, (a key goal of assurance) then the overall impact of AI-driven decisions (whether the decisions are made by a human or are automated) needs to be part of the assurance process. For instance, consider a traffic light control system with a pedestrian crossing. If the metric used to assess performance is average vehicle flow, then the algorithm will perform perfectly if the lights are always on green when a car approaches, but the adverse effect on pedestrians will not be considered. If a system level approach is applied, any metric would consider the existence of pedestrians as well as cars.

One further context to consider is the specific drivers for introducing AI into the system or making it human-free. A key question may be, is the need driven by the industry (e.g. reducing costs) or by the consumers (e.g. improved functionality)? An industry driven AI introduction may require more demonstrations of the system in use for consumers to accept the system and any potential risks. In addition, if there is a clear risk to humans, there is likely to be a greater requirement for increased regulation, and benchmarking is likely to take a different approach compared to cases where the system does not put humans at risk directly. For example, the regulatory requirements ensuring safe interactions between autonomous vehicles and pedestrians will potentially be much stricter than the performance requirements to assess driving in situations that pose a lower risk to human life. The requirements are likely to require the manufacturer to demonstrate both

safe operation of the vehicle and appropriate outcomes in the event of a collision or other undesired event. Whilst such events should be avoided through system design, it is essential to know that the system behaviour will not cause further harm. For instance, if an autonomous car hits a person and keeps on driving there are safety, ethical and legal implications. Consistent assessment throughout the systems life cycle is therefore important to enable a systems view, given the development cycles typically focus on performance metrics while deployment may focus on risk-based metrics in line with regulation.

To support assurance requirements and enable assessment of fully deployed AI systems, carefully defined metrics should be used. Ultimately, when defining metrics we need to take a systems view (the algorithm, its interfaces, and the interactions with the user and other external objects) as well as the wider context for implementation and the problem being solved, to understand what the benchmarking or measurement requirements may be at all stages through the life cycle. A holistic view of a system requires a suite of metrics to capture different aspects of the problem or performance, with the ultimate outcome of the AI use-cases providing the best place to focus for metric development.

3.6 Validation tools and techniques

In addition to requiring a holistic view, metric level information can be difficult to scale or transfer, as is the notion of “how good is good enough?”, supporting the need for a “more than metric” view of AI systems to support confidence building. Confidence is not one-dimensional and reliance on specific indicators only can lead to false or over-confidence in a system. For instance, calibration does not guarantee confidence as the system still must be used in the correct way for the results to be meaningful. The stochastic nature of generative AI and the adaptive and autonomous nature of agentic AI also challenge conventional testing methods and metrics. This leads to the wide-ranging challenges on what tools, methods or techniques should or could be used to support confidence development and decision making. Alongside this, the persuasive capability of the misrepresentation of metrics is key concern, as is the increasing use of quasi or “vibe” based “metrics”. This is a particular challenge area for LLMs, which may be standardised independently from the companies developing them, where there are also gaps around “vibe” and tone of inputs and outputs. With little guidance on which metrics or tools are most appropriate, and no clear approach to feeding information back to developers to

improve their choices, practitioners may end up with sub-optimal or even incorrect information.

There are concerns that poor quality, explainability methods, coupled with model “hallucinations” can lead to false confidence in models and their outputs. Some models are inherently more explainable (e.g. decision trees), but more complex models rely on methods to provide some (possibly subjective) measure of explainability. The rapid development of models and AI systems means that there is consistently a lag with developing and verifying explainability methods, such that models that are in practical use may not be explainable. Some generic tools are available for explainability, though in some cases, use of these tools results in an assumption-based model wrapped around an assumption-based model and does not add to understanding.

In some industries, explainability may not be as important as causality, provided outputs can be verified and risks mitigated. Multi-stage processes may require a chain of logic, e.g. for legal considerations, which may include rare events and assumptions about reliability that are not considered in black box models, or indeed wrapper-models for them. Most notably in imaging data, explanation can lead to over reliance, false confidence and even confirmation bias in models. Input information may also be collected in different ways, e.g. within medical applications uncertainty about a diagnosis may not be quantified but may instead be reflected in the language used in written reports, requiring AI interpretation of textual data to correctly reflect the true level of confidence in the diagnosis. Context specific tools for assessing scenarios such as relevant edge cases or, rare or malicious events are vital, for example in safety critical areas. AI risk classification systems and tools already exist and are being used, but some are not connected to the end use case very well.

Alongside the context from life cycle and end use/outputs, increasingly the heritage of a system (both the algorithm and the training data) is being folded into new evaluation approaches. There is an increasing interest in use of techniques such as transfer learning and domain adaptation more broadly, where algorithms that have been successfully demonstrated in one application are partially retrained or adapted to be deployed in another area. For instance, an image analysis algorithm that has been shown to successfully identify cars could be partially retrained to identify trucks. The heritage, and traceability and provenance more broadly, is therefore increasingly key, possibly more so

than the underlying measurement for establishing trust and therefore confidence. Clear and appropriate handling of assessments of the limitations of a system can be as important as traditional performance indicators to determine what it can't do.

Finally, quantification of uncertainty in AI systems is a growing area in academic research, but this is not yet often seen in industrial situations. Typically, there is interest but a lack of expertise, so for instance ensemble methods may be used but formal uncertainty quantification is not widely used. Unregulated industries are typically less aware of uncertainty quantification than regulated ones which may be a result of the regulatory requirements. Traceability and transparency are important for regulators, where an explanation and rectification are required, and current regulation is very vague. The inability to standardise across applications and outcomes, coupled with the fact that a set of test points and results do not cover the full scale required for validation, means metrics alone as a validation tool cannot cover all needs for users or regulators. Additionally, the selection of metrics used for assessment or validation will vary from company to company depending on their point in the life cycle, their market and their risk appetite, complicating the decision-making process for end users as they cannot compare “apples with apples” by looking at metrics alone. Certification generally assesses the entire process that created a product or system, not just the product, and includes development and deployment. This approach must also apply to AI and requires transparency and comparable evaluation techniques. Assessment of the entire workflow, data and model, is a requirement for fairness when considering the impact on sub-groups and when comparing expectations versus reality in the output results of a model.

Any validation tools or techniques developed, should utilise a framework approach to ensure interpretable testing which takes into consideration heritage, traceability and provenance of any system. Metrics should be folded into these framework approaches with a focus on the ultimate end use case and enable clear and transparent governance over which models are suitable for use in regulated sectors and helping to inform practitioner choices and focus innovation.

4 Directions for the future: uniting the landscape for AI assurance

In terms of AI development, society has come a long way since the National Physical Laboratory 1958 meeting on “Mechanisation of Thought Processes” [1]. AI will become more ubiquitous in society, and with that comes more challenges including legal, ethical, economic and social hurdles to navigate. Looking ahead, AI has the potential to be instrumental in developing a sustainable future. It has clear benefits but also the potential to cause harm if not assessed, managed and regulated properly. This is precisely why confidence and assurance in AI matter, with metrics and measurement fundamental to achieving this. This Bushy House Scientific Meeting was held at a critical time; there are growing regulatory and ethical demands, and the risks of poorly defined AI performance is high. Currently this assurance and regulatory ecosystem is difficult to navigate as it is evolving rapidly, and centralised resources and sources of good practice are still emerging. All stakeholders including developers, 3rd party assurers, regulators and end-users will benefit from clear pathways to test, evaluate, verify and validate AI processes.

This Bushy House Scientific Meeting discussed frameworks and methodologies to enable trustworthiness, assured AI and ultimately providing confidence in AI. Going forward, we recommend the following guiding principles (highlighted in Figure 2 and with more detail below):

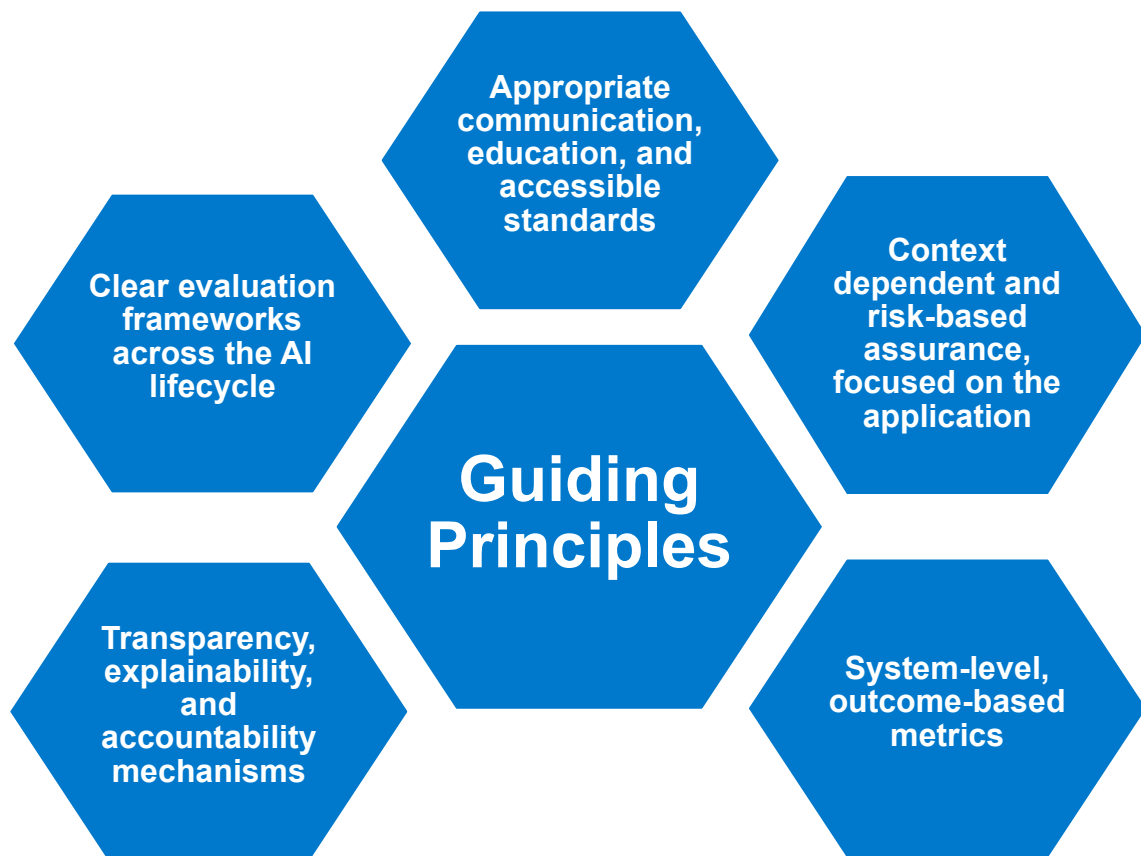


Figure 2: The guiding principles for developing trustworthy AI

Appropriate communication, education, and accessible standards: Effective AI assurance depends on shared understanding among technical experts, decision-makers, regulators, and users, supported by training and standards that are usable by non-specialists. This includes developing a common terminology, and user-friendly tools to enable the continual shared dialogue given this rapidly changing technology.

Context dependent and risk-based assurance, focused on the application: Assurance requirements should vary by use case, risk profile, and sector, with higher thresholds for safety-critical domains and regulation aligned to intended purpose and system impact.

System-level, outcome-based metrics: Benchmarks and metrics should assess real-world outcomes, not just model performance, considering the full socio-technical system; data, models, interfaces, human behaviour, and external context.

Transparency, explainability, and accountability mechanisms: Trust requires interpretable testing, explainable decisions, and clear accountability, the level of which is determined by the application context. These mechanisms can apply across data collection, software development, algorithmic decision making, and human-in-the-loop interactions.

Clear evaluation frameworks across the AI lifecycle: AI assurance must cover both pre-deployment validation and post-deployment continuous monitoring, with traceability, feedback loops, and clarity on responsibilities across developers, deployers, and regulators.

These guiding principles are a call-to-arms in AI assurance. There are many challenges that are sector specific, but this meeting has demonstrated the importance in diverse consultation to build shared understanding, highlighting the cross-sector challenges we face in building confidence in AI. We encourage stakeholders begin thinking about implementing AI assurance now with what tools are available, participate in shaping future standards and metrics through working groups or partnerships and contribute to building educational programs for AI assurance; continued engagement will be vital for long-term success. Building confidence in AI is foundational to realising its societal benefits, with the science of measurement providing the bridge between innovation and deployment, enabling trustworthy AI adoption.

5 References

- [1] "Mechanisation of Thought Processes". The proceedings of the first international AI symposium, November 1958, Teddington, published in 1959 by Her Majesty's Stationery Office, <https://www.historyofinformation.com/detail.php?id=771>
- [2] Citi, web article, "Bank of the Future: the ABCs of Digital Disruption in Finance", 2018. Available at: <https://www.citivelocity.com/citigps/bank-future/>. Retrieved 20/1/26.
- [3] OECD, web resource, [Metrics for Trustworthy AI - OECD.AI](https://oecd.ai/en/catalogue/metrics) Available at <https://oecd.ai/en/catalogue/metrics> . Retrieved 20/1/26.
- [4] Artificial Intelligence Act of the European Union, 2024, Regulation (EU) 2024/1689 of the European Parliament and of the Council, of 13 June 2024, laying down harmonised rules on artificial intelligence, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- [5] AI Standards Hub Standards Database. Available from <https://aistandardshub.org/ai-standards-search/> Retrieved 20/1/26.
- [6] EMN Mathmet Strategic Research Agenda 2023-2033 v1.0, 2023. Available from <https://www.euramet.org/european-metrology-networks/mathmet/strategy/strategic-research-agenda> Retrieved 20/1/26.
- [7] Buchanan, Bonnie, "Artificial intelligence in finance". Alan Turing Institute Research Paper, 2019. Available at: <http://doi.org/10.5281/zenodo.2612537> Retrieved 20/1/26.
- [8] Mitchell, M et al, "Model Cards for Model Reporting", FAT* '19: Conference on Fairness, Accountability, and Transparency, 2019. <https://doi.org/10.1145/3287560.3287596> Retrieved 20/1/26.
- [9] DSIT Guidance document, "Introduction to AI Assurance, 2024. Available from <https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance> Retrieved 20/1/26.

[10] Gichoya J. W et al, "AI recognition of patient race in medical imaging: a modelling study", Lancet Digit Health. 2022 May 11;4(6): e406–e414. doi: 10.1016/S2589-7500(22)00063-2. <https://pmc.ncbi.nlm.nih.gov/articles/PMC9650160/> Retrieved 20/1/26.

6 Appendix - Acknowledgements

This report was written by Louise Wright, Valerie Livina, Jennifer Clarke, Chris Barnes, Spencer Thomas and Paul Duncan based on the discussions held during the Bushy House Scientific Meeting: Getting the Measure of AI (21-22 October 2025).

We thank the reviewers of the document, Dame Wendy Hall, Bonnie Buchanan, Denis Newman-Griffis, Gopal Ramchurn, Ian Gilmore, Vicki Saward, Richard Barker, JT Janssen and Stacie Hoffmann for their insightful comments.

We share the below list as a non-exhaustive list of attendees (in addition to the authors) and wish to acknowledge their contributions to the discussions.

We highlight this report was prepared by NPL and may not reflect the specific views of all individuals present at the meeting.

Name	Organisation
Amanda Niedfeldt	digiLab
Ana Alania	NPL
Andre Burgess	NPL
Ben Harack	Oxford Martin AI Governance Initiative
Bonnie Buchanan	University of Surrey
Damian Ruck	Advai
Danni Croucher	NPL
Denis Newman-Griffis	University of Sheffield Centre for Machine Intelligence
Donnacha Kirk	AI Collaboration Centre
Mostafizur Rahman	The Manufacturing Technology Centre (MTC)
Eliot Gillings	Royal Academy of Engineering
Emma Haynes	NPL
Fanoula Ziouzia	NPL
Franziska Weindauer	TÜV AI.Lab GmbH
Ian Gilmore	NPL
JT Janssen	NPL

Name	Organisation
Professor Jeyan Thiya	STFC-UKRI (Science and Technology Facilities
galingam	Council, UKRI)
Malcolm Skingle	Skingle Solutions Ltd
Mark Thirlwell	British Standards Institution
Matt Holmes	Intellect Frontier
Mike Short	
Natalie Black CBE	Ofcom
Peter Bannister	Romilly Life Sciences
Peter Thompson CBE	NPL
Dame Wendy Hall	University of Southampton
Richard Barker	NPL
Robin Proctor	Royal College of Radiologists
Rod Viggers	UK Research and Innovation
Sascha Eichstädt	Physikalisch-Technische Bundesanstalt (PTB)
Stacie Hoffman	NPL
Sundeep Bhandari	NPL
Thomas Doms	TÜV AUSTRIA Holding AG
Professor Tim Dodwell	digiLab
Vicki Saward	NPL

Getting the measure of AI: building confidence in AI and assurance through metrics and measurement

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**