Department for Science, Innovation & Technology

Introduction to AI assurance

February 2024

Contents

Ministerial foreword	3
1. Executive summary	4
2. AI assurance in context	7
3. The AI assurance toolkit	14
4. AI assurance in practice	27
5. Key actions for organisations	39
6. Additional resources	42

Viscount Camrose

Minister for Artificial Intelligence and Intellectual Property

Artificial Intelligence (AI) is increasingly impacting how we work, live, and engage with others. AI technologies underpin the digital services we use every day and are helping to make our public services more personalised and effective, from improving health services to supporting teachers; and driving scientific breakthroughs so we can tackle climate change and cure disease. However, to fully grasp its potential benefits, AI must be developed and deployed in a safe, responsible way.

The UK government is taking action to ensure that we can reap the benefits of AI while mitigating potential risks and harms. This includes acting to establish the right guardrails for AI through our agile approach to regulation; leading the world on AI safety by establishing the first state-backed organisation focused on advanced AI safety for the public interest; and – since 2021 – encouraging the development of a flourishing AI assurance ecosystem.

As highlighted in our AI regulation white paper in 2023, AI assurance is an important aspect of broader AI governance, and a key pillar of support for organisations to operationalise and implement our five cross-cutting regulatory principles in practice.

AI assurance can help to provide the basis for consumers to trust the products they buy will work as intended; for industry to confidently invest in new products and services; and for regulators to monitor compliance while enabling industry to innovate at pace and manage risk. A thriving AI assurance ecosystem will also become an economic activity in its own right – the UK's cyber security industry, an example of a mature assurance ecosystem, is worth nearly £4 billion to the UK economy.

However, building a mature AI assurance ecosystem will require active and coordinated effort across the economy, and we know that the assurance landscape can be complex and difficult to navigate, particularly for small and medium enterprises. This Introduction to AI assurance is the first in a series of guidance to help organisations upskill on topics around AI assurance and governance. With developments in the regulatory landscape, significant advances in AI capabilities, and increased public awareness of AI, it is more important than ever for organisations to start engaging with the subject of AI assurance and leveraging its critical role in building and maintaining trust in AI technologies.

Department for Science, Innovation & Technology

01

Executive summary

Introduction

The *Introduction to AI assurance* provides a grounding in AI assurance for readers who are unfamiliar with the subject area. This guide introduces key AI assurance concepts and terms and situates them within the wider AI governance landscape. As an introductory guide, this document focuses on the underlying concepts of AI assurance rather than technical detail, however it will include suggestions for further reading for those interested in learning more.

As AI becomes increasingly prevalent across all sectors of the economy, it's essential that we ensure it is well governed. AI governance refers to a range of mechanisms including laws, regulations, policies, institutions, and norms that can all be used to outline processes for making decisions about AI. The goal of these governance measures is to maximise and reap the benefits of AI technologies while mitigating potential risks and harms.

In March 2023, the government published its AI governance framework in a pro-innovation approach to AI regulation. This white paper set out a proportionate, principles-based approach to AI governance, with the framework underpinned by **five cross-sectoral principles**. These principles describe **"what"** outcomes AI systems must achieve, regardless of the sector in which they're deployed. The white paper also sets out a series of tools that can be used to help organisations understand **"how"** to achieve these outcomes in practice: tools for trustworthy AI, including assurance mechanisms and global technical standards.

This guidance aims to provide an accessible introduction to both assurance mechanisms and global technical standards, to help industry and regulators better understand how to build and deploy responsible AI systems. It will be a living, breathing document that we keep updated over time.

The guidance will cover:

- AI assurance in context: Introduction to the background and conceptual underpinnings of AI Assurance.
- The AI assurance toolkit: Introduction to key AI assurance concepts and stakeholders.
- AI assurance in practice: Overview of different AI assurance techniques and how to implement AI assurance within organisations.
- Key actions for organisations: A brief overview of key actions that organisations looking to embed AI assurance can take.

Why is AI assurance important?

Artificial intelligence (AI) offers transformative opportunities for the economy and society. The dramatic development of AI capabilities over recent years, particularly generative AI - including Large Language Models (LLMs) such as ChatGPT - has fuelled significant excitement around the potential applications for, and benefits of, AI systems.

Artificial intelligence has been used to support personalised cancer treatments, mitigate the worst effects of climate change and make transport more efficient. The potential economic benefits from AI are also extremely high. Recent research from McKinsey suggests that generative AI alone could add up to $4.4 trillion to the global economy.

However, there are also concerns about the risks and societal impacts associated with AI. There has been notable debate about the potential existential risks to humanity but there are also significant, and more immediate, concerns relating to risks such as bias, a loss of privacy and socio-economic impacts such as job losses.

When ensuring the effective deployment of AI systems many organisations recognise that, to unlock the potential of AI systems, they will need to secure public trust and acceptance. This will require a multi-disciplinary and socio-technical approach to ensure that human values and ethical considerations are built-in throughout the AI development lifecycle.

AI assurance is consequently a crucial component of wider organisational risk management frameworks for developing, procuring, and deploying AI systems, as well as demonstrating compliance with existing - and any relevant future - regulation. With developments in the regulatory landscape, significant advances in AI capabilities and increased public awareness of AI, it is more important than ever for organisations to start engaging with AI assurance.

Department for Science, Innovation & Technology

02

AI assurance in context

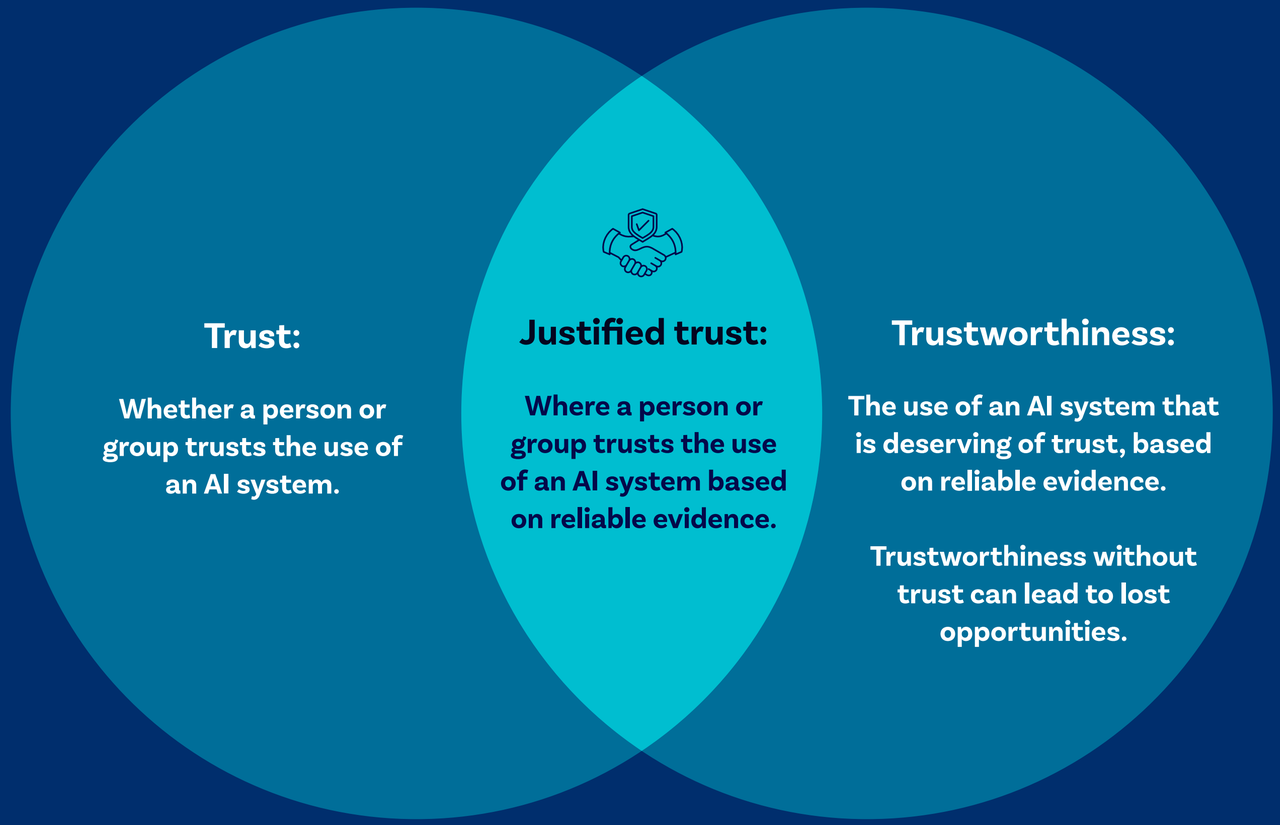
The importance of trust

The term ‘assurance’ originally derived from accountancy but has since been adapted to cover areas including cyber security and quality management. Assurance is the process of measuring, evaluating and communicating something about a system or process, documentation, a product or an organisation. In the case of AI, assurance measures, evaluates and communicates the trustworthiness of AI systems. When developing and deploying AI systems, many organisations recognise that to unlock their potential, a range of actors – from internal teams to regulators to frontline users – will need to understand whether AI systems are trustworthy. Without trust in these systems, organisations may be less willing to adopt AI technologies because they don’t have the confidence that an AI system will actually work or benefit them.

They also might not adopt AI for fear of facing reputational damage or public backlash. Without trust, consumers will also be cautious about using these technologies. Although awareness of AI is very high amongst the public and has increased over the last year, their primary associations with AI typically reference uncertainty.

AI assurance processes can help to build confidence in AI systems by measuring and evaluating reliable, standardised, and accessible evidence about the capabilities of these systems. It measures whether they will work as intended, hold limitations, and pose potential risks, as well as how those risks are being mitigated to ensure that ethical considerations are built-in throughout the AI development lifecycle.

The relationship between trust, trustworthiness and justified trust



Justified trust

By building **trust in AI systems** through effective communication to appropriate stakeholders, and ensuring the **trustworthiness of** AI systems, AI assurance will play a crucial role in enabling the responsible development and deployment of AI, unlocking both the economic and social benefits of AI systems.

AI assurance and governance

In March 2023, the UK government outlined its approach to AI governance through its white paper, a pro-innovation approach to AI regulation, which set out the key elements of the UK's proportionate and adaptable regulatory framework. It includes **five cross-sectoral principles** to guide and inform the responsible development and use of AI in all sectors of the economy:

Safety, Security and Robustness

AI systems should function in a robust, secure and safe way, and risks should be continually identified, assessed and managed.

Appropriate Transparency and Explainability

AI systems should be appropriately transparent and explainable.

Fairness

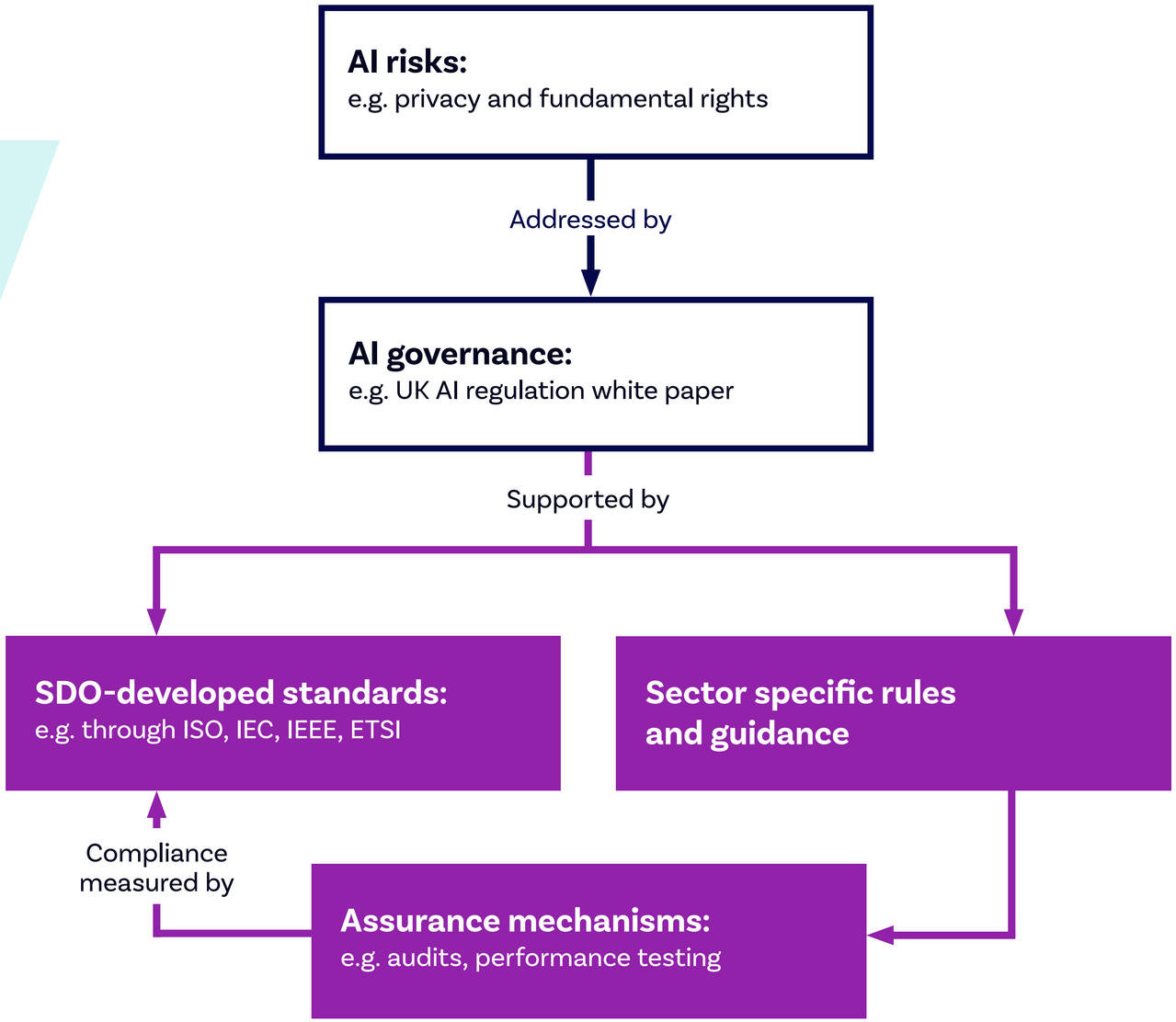
AI systems should not undermine the legal rights of individuals or organisations, discriminate unfairly against individuals, or create unfair market outcomes.

Accountability and Governance

Governance measures should be in place to ensure effective oversight of the supply of AI systems, with clear lines of accountability across the AI life cycle.

Contestability and Redress

Where appropriate, users, affected third parties and actors in the AI lifecycle should be able to contest an AI decision or outcome that is harmful or creates material risk of harm.



UK regulatory framework

AI assurance will play a critical role in the implementation and operationalisation of these principles. The principles identify specific goals – the **"what"** - that AI systems should achieve, regardless of the sector in which they are deployed. AI assurance techniques and standards (commonly referred to as "tools for trustworthy AI") can support industry and regulators to understand **"how"** to operationalise these principles in practice, by providing agreed-upon processes, metrics, and frameworks to support them to achieve these goals.

AI governance and regulation

Due to the unique challenges and opportunities raised by AI in particular contexts, the UK's approach to AI governance focuses on **outcomes** rather than the technology itself – acknowledging that potential risks posed by AI will depend on the context of its application. To deliver this outcomes-based approach, existing regulators will be responsible for interpreting and implementing the regulatory principles in their respective sectors and establishing clear guidelines on how to achieve these outcomes within a particular sector. By outlining processes for making and assessing verifiable claims to which organisations can be held accountable, AI assurance is a key aspect of broader AI governance and regulation.

Through AI assurance, organisations can measure whether systems are trustworthy and demonstrate this to government, regulators, and the market. They can also gain a competitive advantage, building customer trust and managing reputational risk. On one hand, using assurance techniques to evaluate AI systems can build trust in consumer-facing AI systems by demonstrating adherence to the principles of responsible AI (fairness, transparency etc.) and/or relevant regulation/legislation. On the other hand, using assurance techniques can also help identify and mitigate AI-related risks to manage reputational risks and avoid negative publicity. This helps to mitigate greater commercial risks, in which high-profile failures could lead to reduced customer trust and adoption of AI systems.

Outside of the UK context, supporting cross-border trade in AI will also require a well-developed ecosystem of AI assurance approaches, tools, systems, and technical standards which ensure international interoperability between differing regulatory regimes. UK firms need to demonstrate risk management and compliance in ways that are understood by trading partners and consumers in other jurisdictions.

Spotlight: AI assurance and frontier AI

AI assurance, the process of measuring, evaluating and communicating the trustworthiness of AI systems, is also relevant at the "frontier" of AI.

The Bletchley Declaration, signed by countries that attended the November 2023 UK AI Safety Summit, recommended that firms implement assurance measures. This includes safety testing, evaluations, and accountability and transparency mechanisms to measure, monitor and mitigate potentially harmful capabilities of frontier AI models.

The UK's AI Safety Institute (AISI) – the first state-backed organisation focused on advanced AI safety for the public interest – is developing the sociotechnical infrastructure needed to identify potential risks posed by advanced AI. It will offer novel tools and systems to mitigate these risks, and support wider governance and regulation, further expanding the AI assurance ecosystem in the UK.

Department for Science, Innovation & Technology

03

The AI assurance toolkit

Measure, evaluate and communicate

Assurance requires robust techniques that help organisations to **measure and evaluate** their systems and **communicate** that their systems are trustworthy and aligned with relevant regulatory principles. More detail on this is provided to the right.

Given the complexity of AI systems, we require a toolbox of different products, services, and standards to assure them effectively.

1. Measure

Gathering qualitative and quantitative data on how an AI system functions, to ensure that it performs as intended. This might include information about performance, functionality, and potential impacts in different contexts. Additionally, you may need to ensure you have access to documentation about the system design and any management processes to ensure you can evaluate effectively.

2. Evaluate

Activities encompassing techniques to assess the risks and impacts of AI systems and inform further decision-making. This might include evaluating the implications of an AI system against agreed benchmarks set out in standards and regulatory guidelines to identify issues.

3. Communicate

A range of communication techniques can be applied to ensure effective communication both within an organisation and externally. This might include collating findings into reports/presenting information in a dashboard as well as external communication to the public to set out steps an organisation has taken to assure their AI systems. In the long-term this may include activities like certification.

AI assurance mechanisms

There is a spectrum of AI assurance mechanisms that can, and should, be used in combination with one another across the AI lifecycle. These range from qualitative assessments which can be used where there is a high degree of uncertainty, ambiguity and subjectivity, for example thinking about the potential risks and societal impacts of systems, to quantitative assessments for subjects that can be measured objectively and with a high degree of certainty, such as how well a system performs against a specific metric, or if it conforms with a particular legal requirement. The table on the right details a sample of some key assurance techniques that organisations should consider as part of the development and/or deployment of AI systems.

Risk assessment: Used to consider and identify a range of potential risks that might arise from the development and/or deployment of an AI product/ system. These include bias, data protection and privacy risks, risks arising from the use of a technology (for example the use of a technology for misinformation or other malicious purposes) and reputational risk to the organisation.	**(Algorithmic) impact assessment:** Used to anticipate the wider effects of a system/product on the environment, equality, human rights, data protection, or other outcomes.	**Bias audit:** Assesses the inputs and outputs of algorithmic systems to determine if there is unfair bias in the input data, the outcome of a decision or classification made by the system.
Compliance audit: Involves reviewing adherence to internal policies, external regulations and, where relevant, legal requirements.	**Conformity assessment:** The process of conformity assessment demonstrates whether a product or system meets relevant requirements, prior to being placed on the market. Often includes performance testing.	**Formal verification:** Formal verification establishes whether a system satisfies specific requirements, often using formal mathematical methods and proofs.

In addition, a key baseline requirement of AI assurance is ensuring that your data and systems are safe and secure. The National Cyber Security Centre (NCSC) has developed a range of resources and courses, including the Cyber Essentials certification to help organisations develop their cyber security capabilities.

There will never be a silver bullet for AI assurance. Rather, multiple assurance techniques will need to be used in combination with one another across the lifecycle. It is therefore an important challenge for organisations to ensure that suitable assurance techniques and mechanisms are adopted, depending on the context in which a system is being deployed. However, this also allows for a proportionate approach to assurance, with low-risk use-cases able to rely on a smaller range of assurance techniques, and high-risk use-cases utilising a more robust combination of assurance techniques.

AI assurance and standards

To provide a consistent baseline, and increase their effectiveness and impact, AI assurance mechanisms should also be underpinned by available **global technical standards**. These are consensus-based standards developed by global standards development organisations (SDOs) such as the International Standards Organisation (ISO). Global technical standards are essentially agreed ways of doing things, designed to allow for shared and reliable expectations about a product, process, system or service. Global technical standards allow assurance users to trust the evidence and conclusions presented by assurance providers – without standards we have advice, not assurance. There are different kinds of standards that can support a range of assurance techniques. These include:

Foundational and terminological:

Provide shared vocabularies, terms, descriptions and definitions to build common understanding between stakeholders.

Interface and architecture:

Define common protocols, formats and interfaces of a system, for example interoperability, infrastructure, architecture and data management standards.

Measurement and test methods:

Provide methods and metrics for evaluating properties (e.g., security, safety) of AI systems.

Process, management, and governance:

Set out clear processes and approaches for best practice in organisational management, governance and internal controls.

Product and performance requirements:

Set specific criteria and thresholds to ensure that products and services meet defined benchmarks, safeguarding consumers by setting safety and performance requirements.

Standards for bias audit

Bias Audit

Assessing the inputs and outputs of algorithmic systems to determine if there is unfair bias in the input data, the outcome of a decision or classification made by the system.

Foundational and terminological standards What do we mean by “bias” and “fairness” in this context? What are we trying to measure?	**IEEE P7003** **ISO/IEC TR 24027**
Process, management and governance standards What organisational and governance processes do you have in place to support responsible and fair innovation?	**ISO/IEC 42001** **ISO/IEC 23894**
Measurement and test methods What are the methods and metrics you’re using to measure bias in your AI system?	**ISO/IEC TR 24027** **ISO/IEC TS 12791**
Product and performance requirements What is an acceptable output of my bias audit, in order for me to safely deploy this system? Is some bias acceptable?	**ISO/IEC TR 24027** **ISO/IEC 12791**

All these types of standards can underpin and help to enable a range of assurance techniques. The visual to the left showcases an example of how each of these standard types may be relevant when conducting a bias audit.

*This is not an exhaustive list of standards. Rather, these are an indicative example of the types of standards that are available.

Spotlight: AI Standards Hub

The AI Standards Hub is a joint initiative led by The Alan Turing Institute, the British Standards Institution (BSI), and the National Physical Laboratory (NPL), supported by the government. The Hub's mission is to advance trustworthy and responsible AI with a focus on the role that global technical standards can play as governance tools and innovation mechanisms. The AI Standards Hub aims to help stakeholders navigate and actively participate in global AI standardisation efforts and champion global technical standards for AI.

Dedicated to knowledge sharing, community and capacity building, and strategic research, the hub seeks to bring together industry, government, regulators, consumers, civil society and academia with a view to:

- Increasing awareness and contributions to global technical AI standards in line with UK values.
- Increasing multi-stakeholder involvement in AI standards development.
- Bringing the UK AI community together to encourage more coordinated engagement in global AI technical standards development.
- Increasing research and analysis of global AI technical standards, including with international partners, to ensure standards are shaped in line with our shared values.

To learn more, visit the AI Standards Hub website.

The AI assurance ecosystem

There is a growing market of AI assurance providers who supply the assurance systems and services required by organisations who either don't have in-house teams offering internal assurance capabilities, or who require additional capabilities on top of those they have internally. As with assurance techniques and mechanisms, there is no single 'type' of assurance provider, with some third-party providers offering specific technical tools, whilst others offer holistic AI governance platforms. There are also diversified professional services firms who offer assurance 'as a service', supporting clients to embed good governance and assurance practices. Due to its relationship with wider organisational risk management, AI assurance is often seen as one part of an organisation's Environmental, Social and Corporate Governance processes (ESG).

However, AI assurance isn't just limited to a selection of mechanisms and standards and the assurance teams and providers that use them. A range of actors need to check that AI systems are trustworthy and compliant, and to communicate evidence of this to others. These actors can each play several interdependent roles within an assurance ecosystem. The next few pages provide examples of key supporting stakeholders and their role within the AI assurance ecosystem.

Regulators

Role:
To set regulation and best practice in their relevant domains and (where required) encourage, test and verify that AI systems are compliant with their regulations. Regulators will also incentivise best practice and create the conditions for the trustworthy development and use of AI.

Example:
Individual regulators will be responsible for developing regulatory guidance and oversight for the deployment of technology in their respective areas.

The ICO has already developed regulatory guidance and toolkits relating to how data protection regulation applies to AI.

Regulators are also implementing a range of sandboxes and prize challenges to support regulatory innovation in AI.

Audience:

- Affected individuals
- Organisations developing AI systems
- Organisations procuring AI systems

Accreditation bodies

Role:
To attest to the ongoing competence, impartiality of AI services provided by third party assurance providers against international standards. This will build trust in auditors, assessors and suppliers throughout the AI assurance ecosystem.

N.B. Accreditation bodies will not be able to accredit organisations offering 'assurance as a service'. The United Kingdom Accreditation Service (UKAS) cannot review organisations offering services to third-party clients whose services may also be reviewed by UKAS, as it would represent a conflict of interest.

Example:
UKAS is the UK's sole national accreditation body. It is appointed by the government to assess a third-party organisation known as a conformity assessment bodies who provide a range of services including AI assurance.

Services offered by UKAS include certification, testing, inspection and calibration.

Please review the UK government's policy on accreditation and conformity assessments if you want more information.

Audience:

- Affected individuals
- Government
- Regulators
- Organisations developing AI systems
- Organisations procuring AI systems
- Other assurance providers

Government

Role:
Drive the development of an AI assurance ecosystem that supports compliance with laws and regulations, in a way that does not hinder economic growth.

Example:
The DSIT AI assurance programme supports the development of a robust and sustainable AI assurance ecosystem.

Work to date includes building knowledge of the AI assurance market, its drivers and barriers, highlighting emerging assurance techniques (through the Portfolio of Assurance Techniques) and supporting the development of novel assurance techniques, for example through the Fairness Innovation Challenge.

Audience:

- **Affected individuals**
- **Regulators**
- **Organisations developing AI systems**
- **Organisations procuring AI systems**

Standards bodies

Role:
To convene actors including industry and academia to develop commonly accepted standards that can be evaluated against.

Example:
There are both international and national standards bodies. International standards bodies include the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC) as well as national standards group.

The British Standards Institute (BSI) is the UK's national standards body. BSI represents UK stakeholders at specific regional and international standards bodies that are part of the standards system. BSI along with the National Physical Laboratory (NPL) and the UK accreditation service (UKAS) make up the UK's national quality infrastructure for standards (see here for more information).

Audience:

- **Regulators**
- **Organisations developing AI systems**
- **Organisations procuring AI systems**

Research bodies

Role:
Contribute to research on potential risks or develop leading-edge assurance systems.

Example:
The Alan Turing Institute is the national institute for data science and AI, working to advance research and apply it to national and global challenges – including via a research programme on AI assurance.

The Alan Turing Institute is currently collaborating with the University of York on the Assuring Autonomy International Programme to build on and harmonise existing research into trustworthy and ethical assurance. This project includes the development of open and reproducible tools to help project teams meet ethical and regulatory best practices in health research and healthcare for a range of data-driven technologies.

Audience:

Government | Regulators | Other assurance providers

Civil society organisations

Role:
Through oversight and stakeholder convening, civil society organisations can support multi-stakeholder feedback and scrutiny on AI systems. They can also keep the public/industry informed of emerging risks and trends through external advocacy and develop assurance thought leadership and best practice.

Example:
Civil society organisations are working on developing resources and templates to support AI assurance processes. For example, the Ada Lovelace Institute's Algorithmic Impact Assessment in Healthcare project has developed a template algorithmic impact assessment (AIA) in a healthcare context. This aims to ensure that algorithms that use public sector data are evaluated and governed to produce benefits for society, governments, public bodies and technology developers, as well as the people represented in the data and affected by the technologies and their outcomes.

Audience:

Affected individuals | Government | Regulators | Organisations developing AI systems | Organisations procuring AI systems

Professional bodies

Role:
To define, support, and improve the professionalisation of assurance standards and to promote information sharing, training, and good practice for professionals, which can be important both for developers and assurance service providers.

Example:
There are not currently any professional bodies with Chartered Status, with a focus on AI assurance.

However, the UK Cyber Security Council has recently been created and made responsible for standards of practice for cyber security professionals. This model that could be adopted for AI assurance in the future.

The International Association of Algorithmic Auditors (IAAA) is another recently formed body, hoping to professionalise AI auditing by creating a code of conduct for AI auditors, training curriculums, and eventually, a certification programme.

Audience:

Regulators

Organisations developing AI systems

Organisations procuring AI systems

Spotlight: Fairness Innovation Challenge

A range of the above stakeholders are already working together to grow the UK's AI assurance ecosystem through the Fairness Innovation Challenge. The Challenge, run by DSIT in partnership with Innovate UK, brings together government, regulators, academia, and the private sector to drive the development of novel socio-technical approaches to fairness and bias audit – a currently underdeveloped area of research, with most bias audits measuring purely technical or statistical notions of fairness. The Challenge will provide greater clarity about how different assurance techniques can be applied in practice, and work to ensure that different strategies to address bias and discrimination in AI systems comply with relevant regulation, include data protection and equalities law.

Spotlight: IAPP Governance Center

The International Association of Privacy Professionals (IAPP) is the largest global information privacy community and resource centre, with more than 80,000 members. It is a non-profit, policy-neutral professional association helping practitioners develop their capabilities and organisations to manage and protect their data.

In Spring 2023, the IAPP AI Governance Center was launched, in recognition of the need for professionals to establish the trust and safety measures that will ensure AI fulfils its potential to serve society in positive and productive ways.

Through the AI Governance Center, the IAPP provides professionals tasked with AI governance, risk, and compliance with the content, resources, networking, training and certification they need to manage the complex risks of AI. This allows AI governance professionals to share best practices, track trends, advance AI governance management issues, standardise practices and access the latest educational resources and guidance.

In December 2023, the IAPP AI Governance Center published a report, based on survey responses from over 500 individuals from around the world, on organisational governance issues relating to the professionalisation of AI governance. The report covered the use of AI within organisations, AI governance as a strategic priority, the AI governance function within organisations, the benefits of AI-enabled compliance, and AI governance implementation challenges. Notably, IAPP research has found that:

60%

of respondents indicated their organisation has already established a dedicated AI governance function or is likely to in the next 12 months.

31%

cited a complete lack of qualified AI governance professionals as a key challenge.

56%

indicated they believe their organization does not understand the benefits and risks of AI deployment.

To learn more visit the IAPP AI Governance Center.

Department for Science, Innovation & Technology

04

AI assurance in practice

AI assurance landscape

There are a wide range of techniques and subjects in scope for AI assurance, with considerable variations in metrics and methodologies across sectors, situations and systems. The diagram to the right demonstrates how this can look in practice at a high-level. DSIT shortly plans to publish additional sector-specific guidance to provide more detail about AI assurance in particular contexts.

My organisation wants to demonstrate it has...

	Good internal governance processes around AI	Understood the potential risks of AI systems it is buying	Made sure AI systems it is building or buying adhere to existing regulations for data protection
Assurance mechanism	Conformity assessment	Algorithmic impact assessment	Compliance audit
Provider	UKAS accredited conformity assessment body	UKAS accredited conformity assessment body	Third party assurance provider
Measured against	SDO-developed standards, e.g. ISO/IEC 42001, AI Management System	(Self) assessment against proprietary framework or responsible AI toolkit	UK GDPR

AI assurance spectrum

Given the variety of use-cases and contexts that AI assurance techniques can be deployed in and the need for assurance across the **AI development lifecycle**, a broad range of assurance techniques beyond **purely technical solutions** will need to be applied.

Within scope of AI assurance techniques are:

Training data:

Training data is the information an AI system uses to 'learn' from during its initial development. Additional training data may be used later in the lifecycle to update the model.

AI models:

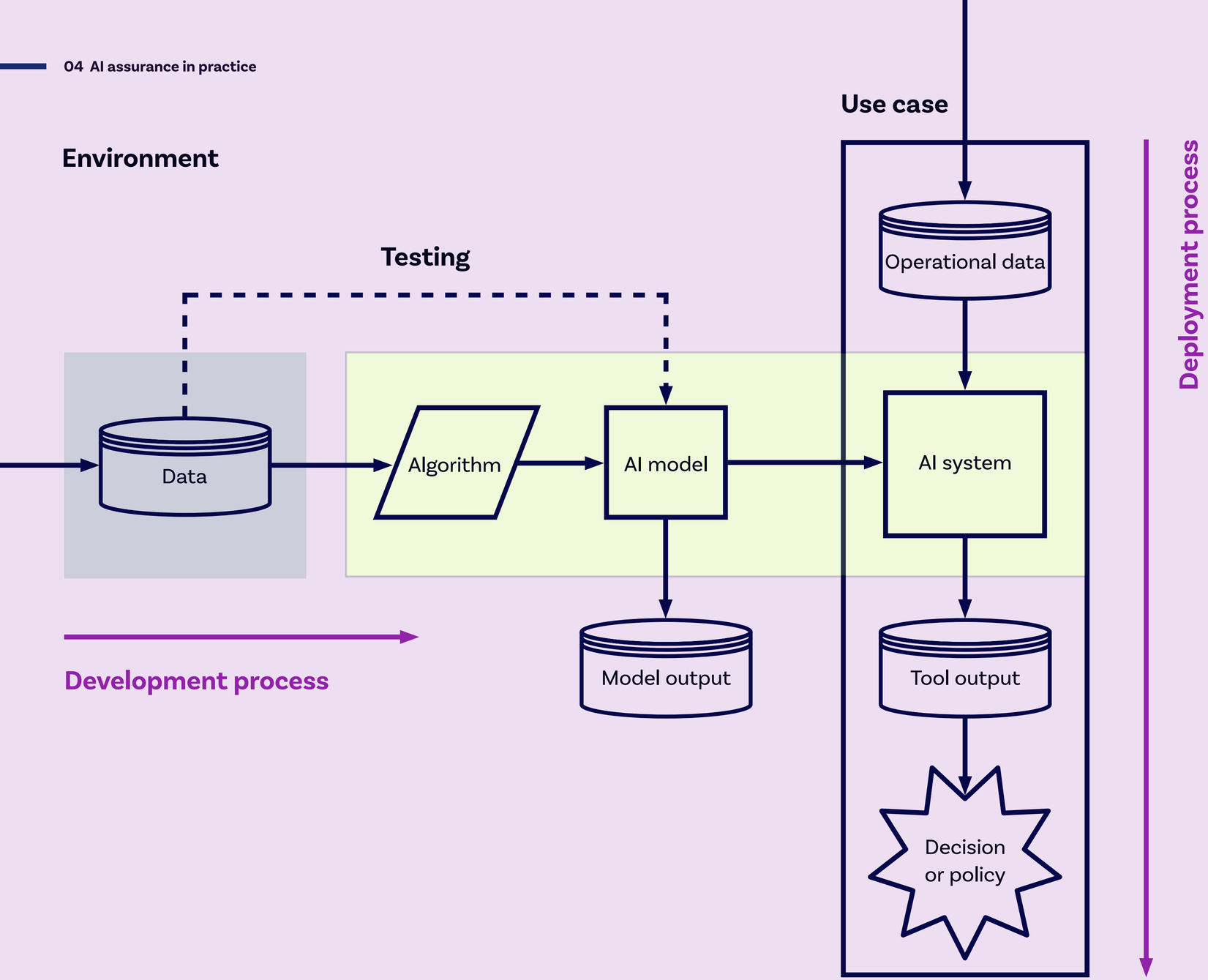
AI models are artefacts of machine learning methods, where the model 'learns' how to solve certain tasks by identifying patterns in the training data. They are constructed to help analyse, explain, predict, or control the properties of real-world systems. GPT-4 is an example of an AI model.

AI systems:

AI systems are the products, tools, applications or devices that utilise AI models to help solve problems. They may comprise a single model, or multiple models. AI systems are the operational interfaces to AI models – they incorporate the technical structures and processes that allow models to be used by non-technologists. ChatGPT is an example of an AI system, which uses GPT-4.

Broader operational context:

This refers to how AI systems are deployed within the wider organisational context – the broader systems that support decision making and the delivery of outcomes. This can include the impact of a system on groups and individuals and the balance of liabilities between users and organisations.



The diagram to the left demonstrates how the development and deployment lifecycles fit together. The horizontal development process is relevant where you are building AI systems in-house, and the vertical deployment process is relevant where you are deploying procured or in-house AI systems.

Assurance of relevant data, the model and/or the system, and the governance in place around the process is relevant in both situations.

Assuring data, models, systems and governance in practice

Data

Data is the foundation of AI. Without the collection of high-quality, robust, and ethically sourced data, AI technologies will be unable to operate effectively or maintain trust. Organisations should put in place robust and standardised processes for handling data, including:

- An effective organisational data strategy;
- Clear employee accountabilities for data and;
- Robust, standardised and transparent processes for data collection, processing and sharing.

All organisations processing data must comply with existing legal requirements, in particular, UK GDPR and the Data Protection Act 2018. All systems using personal data must carry-out a data protection impact assessment (DPIA).

A DPIA should consider risks and impacts to individuals, the rights and freedoms of individuals, groups and society and map-out planned data processing activities.

Models and Systems

Assurance techniques and practices designed to improve justified trust in AI models and systems will work to ensure that they function as intended, and that they produce beneficial outcomes. This includes ensuring that outputs are accurate, and minimising potential harmful outcomes such as unwanted bias. A range of assurance tools and techniques can be used to evaluate AI models and systems, including impact assessments, bias audits and performance testing.

When setting metrics, teams should build in rigorous software testing and performance assessment methodologies with comparisons to clear performance benchmarks.

Governance

As a foundation, all organisations should integrate robust **organisational governance** frameworks for AI systems. There are core steps organisations should build into governance processes to enable the effective **evaluation** and **measurement** of risks and biases associated with AI and to support clear and accurate communication to ensure concerns and issues are flagged.

The next few pages lay out some examples of the assurance mechanisms we introduced previously, highlighting how they may be used to assure AI systems, and what outcomes an organisation could expect from such mechanisms being used in practice. Further examples of real-world AI Assurance techniques developed by assurance providers can be found in DSIT's Portfolio of Assurance Techniques.

As a baseline, core governance processes include:

- Clear, standardised internal transparency and reporting processes and lines of responsibility, with a named person responsible for data management and clear governance and accountability milestones built into the project design.
- Clear avenues for escalation and staff (at all levels) empowered to flag concerns to appropriate levels.
- Clear processes to identify, manage and mitigate risks.
- Quality assurance processes built-in throughout the AI lifecycle.
- External transparency and reporting processes.
- Ensuring the right skills and capabilities are in place and that AI assurance capability and infrastructure is adequately funded.

Mitigations to identified risks and issues can include:

- Internal mitigations, such as metrics designed to identify, and measures to correct, poor-performing or unfair algorithms, and amendments to internal governance processes to improve risk management.
- External (public) redress for affected individuals such as through a clearly communicated appeals process. Such mechanisms should be simple and – importantly – effective.

Risk assessment

Risk assessments are used to consider and identify a range of potential risks that might arise from the development and/or deployment of an AI product/systems.

Models

Tools

Background

An online education platform is exploring ways of using AI to personalise video content presented to users. The company conducts a risk assessment to explore potential costs and benefits, including effects on reputation, safety, revenue, and users.

Process

Staff are encouraged to share their thoughts and concerns in workshops designed to capture and classify a range of potential risks. The workshop is followed by more detailed questionnaires. The company's internal audit staff then assess the risks to quantify and benchmark them and present their findings in an internal report to inform future decision-making.

Outcomes

The risk assessment results in a clearer and shared appreciation/understanding of potential risks, which is used to create mitigation strategies and build resilience to manage future change across the organisation.

Impact assessment

Impact assessments are used to anticipate the wider effects of a system/product on the environment, equality, human rights, data protection, or other outcomes.

Models | Data | Tools | Governance

Background

A specialist waste collection company is conducting an Impact Assessment before deploying a new AI system to manage its waste sorting system more efficiently.

Process

The company uses a tool to assess different aspects of the potential outcomes of the new AI system. During the planning phase, the organisation has adapted a toolkit - including a detailed questionnaire - from a publicly available resource. The toolkit includes categories that cover environmental impacts, fairness, transparency and inclusivity and guides assessors to map out the rationale for deploying the new system, as well as potential consequences. Results are published in a report alongside mitigations and reflections on the findings.

Outcomes

The impact assessment enables the company to develop a clearer understanding of the potential opportunities and costs from deploying a new AI system. The final report also provides a record of actions that could help mitigate harmful outcomes.

Bias audit

Bias audits focus on assessing the inputs and outputs of algorithmic systems to determine whether there is unfair bias in the outcome of a decision, classification made by the system, or input data.

Models **Data**

Tools **Governance**

Background

A growing company is using an AI system they have built and trained to sift candidates in a recruitment campaign. This is to ensure that their approach does not introduce unintended biases to the system.

Process

A third-party auditor works with internal staff to define the scope and objectives for the audit. It then requests data from the company and access to their systems, to undertake a technical audit to measure outcomes from the sifting algorithms to identify potential bias.

Outcomes

The audit reveals that the model is more likely to pass candidates who live in a certain area. This might introduce biases as home addresses may correlate with other protected characteristics e.g. ethnicity/age/disability.

A third-party auditor collaborates with system developers to analyse findings and present recommendations. They also support the developers to implement mitigations and ensure that systems are designed in a way that minimises unfair bias.

Compliance audit

A compliance audit involves reviewing adherence to internal policies, external regulations and, where relevant, legal requirements.

Governance

Background

A UK-based company that uses AI in its manufacturing processes is undertaking work to ensure its products comply with regulatory requirements overseas. The company is planning to expand to sell a new product in international markets.

Process

The company starts with a compliance audit. They work with a third-party service provider to review the extent to which their products and processes adhere to internal policies and values, as well as with regulatory and legal requirements.

The company uses a governance platform, purchased from a specialist third-party provider of AI assurance systems. This includes features that will update, inform and guide developers of relevant requirements that apply across different markets.

Outcomes

After applying these assurance techniques, the company has a clearer understanding of how well it complies with requirements and standards across different markets and regions.

Conformity assessment

Conformity assessment activities that are performed by a person or organisation that is independent of the product, process, system, claim etc and has no user interest in that object. Used to demonstrate that a product, process, system, claim etc conforms with specified requirements that are defined in normative documents e.g. regulations, standards and/or technical specifications.

Conformity assessment may include activities such as testing, inspection, validation, verification and certification. Contingent on the level of risk, conformity assessment activities should be undertaken by an independent third-party conformity assessment body. It is UK government policy that where third-party conformity assessment services are sought, they should be obtained from an organisation accredited by UKAS.

Models

Tools

Background

A large technology company is using a UKAS accredited product certification body to demonstrate that its products conform with applicable product standards, initially, and on an ongoing basis to give customers/ regulators confidence that normative requirements are being continuously achieved and maintained.

Process

An accredited certification body has developed a scheme that applies to the products produced by the technology company. The requirements of the scheme include (but are not limited to) the methodology to be used for performing the conformity assessment activities. The product scheme incorporates requirements, in addition to those specific to the products, that relate to the consistent operation of a management system to give confidence in the ongoing conformity of production.

Outcomes

The resulting product certificate(s) demonstrate that the products produced by the technology company continue to conform with the requirements of the certification scheme (for as long as the certificate(s) remain valid).

Formal verification

Formal verification establishes whether a system satisfies specific requirements, often using formal mathematical methods and proofs.

Models

Tools

Background

A bank is using formal verification to test a newly updated AI model that will support the assessment of mortgage applications to ensure the models are robust and any risks associated with its use are verified.

Process

The bank is working with a consultancy to provide a robust assessment of its software prior to deployment. The use of a third-party assurance provider helps to ensure that the assessment is impartial and allows the bank to assess its algorithms thoroughly by making use of specialist expertise. The verification process uses formal mathematical methods to assess whether the system updates satisfy key requirements. Following the assessment, the consultants present results in a detailed report, which highlights any errors or risk factors the assessment has flagged.

Outcomes

The thorough scrutiny of the formal verification process ensures that any potential risks or errors are identified before the system is in use. This is particularly important in financial services, where errors could have severe consequences for users, the bank's reputation and ability to meet regulation. The results also provide an objective and quantifiable measurement of the model's functionality. This enhances security, and confidence from users and shareholders.

Department for Science, Innovation & Technology

05

Key actions for organisations

Steps to build AI assurance

AI assurance is not a silver bullet for responsible and ethical AI, and whilst the ecosystem is still developing there remain limitations to, and challenges for, successfully assuring AI systems. However, early engagement and proactive consideration of likely future governance needs, skills and/or technical requirements can help to build your organisation's assurance capabilities.

If you are an organisation interested in further developing your AI assurance understanding and capability, you may want to consider the following steps:

1.

Consider existing regulations

While there is not currently statutory AI regulation in the UK, there are existing regulations that are relevant for AI systems. For example, systems must adhere to existing regulation such as UK GDPR, the Equality Act 2010 and other industry-specific regulation.

2.

Upskill within your organisation

Even whilst the ecosystem is still developing, organisations should be looking to develop their understanding of AI assurance and anticipating likely future requirements. The Alan Turing Institute has produced several training workbooks focused on the application of AI governance in practice, and the UK AI Standards Hub has a training platform, with e-learning modules on AI Assurance.

3.

Review internal governance and risk management

Effective AI assurance is always underpinned by effective internal governance processes. It's crucial to consider how your internal governance processes ensure risks and issues can be quickly escalated, and effective decision-making can be taken at an appropriate level. The US's National Institute for Science and Technology (NIST) has developed an in-depth Risk Management Framework (RMF) that can support the management of organisational risk.

4.

Look out for new regulatory guidance

Over the coming years, regulators will be developing sector-specific guidance setting out how to operationalise and implement the proposed regulatory principles in each regulatory domain. For example, the ICO has developed guidance on AI and data protection for those in compliance-focused roles. The UK government has also published initial guidance to regulators as part of its response to the AI regulatory white paper consultation.

5.

Consider involvement in AI standardisation

Private sector engagement with SDOs is crucial for ensuring the development of robust and universally accepted standards protocols, particularly from SMEs who are currently underrepresented. Consider engaging with standards bodies such as BSI. Visit the AI Standards Hub for information and support for implementing AI standards.

If you'd like more information about AI assurance and how it can be applied to your own organisation, don't hesitate to get in contact with the AI assurance team at: **ai-assurance@dsit.gov.uk**

Department for Science, Innovation & Technology

06

Additional resources

Additional resources for those interested in understanding more about AI assurance

Department for Science, Innovation and Technology (2023): A pro-innovation approach to AI regulation

Department for Science, Innovation and Technology (2023): Accreditation and Certification: Six Lessons for an AI Assurance Profession from Other Domains

Department for Science, Innovation and Technology (2022): Industry Temperature Check: Barriers and Enablers to AI Assurance

Department for Science, Innovation and Technology (2023): Portfolio of Assurance Techniques

UK AI Standards Hub: Upcoming Events

UK AI Standards Hub: AI Standards Database

Burr, C., & Leslie, D. (2022). Ethical assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies

National Institute of Science and Technology (NIST): AI Risk Management Framework

National Cyber Security Centre (NCSC): Cyber Essentials

Department for
Science, Innovation
& Technology

magnetic