techUK
FOR WHAT COMES NEXT

Ethics in Action: From White Paper to Workplace

How Practical Implementation of the UK AI White Paper's Ethics Principles can be Achieved Through Assurance and Standard

November 2024

About techUK

techUK is a membership organisation launched in 2013 to champion the technology sector and prepare and empower the UK for what comes next, delivering a better future for people, society, the economy and the planet. It is the UK's leading technology membership organisation, with around 1,000 members spread across the UK. We are a network that enables our members to learn from each other and grow in a way which contributes to the country both socially and economically. By working collaboratively with government and others, we provide expert guidance and insight for our members and stakeholders about how to prepare for the future, anticipate change and realise the positive potential of technology in a fast-moving world.

About techUK's Digital Ethics Programme

In an increasingly digital world, it's important that technology is used to improve and enhance the quality of people's everyday lives. Embedding ethical principles, such as transparency, accountability and explainability, into the creation of products, tools and services is essential for building public trust and confidence in technology. techUK focuses on resolving some of the most difficult ethical challenges, to ensure tech works for people and responsible innovation can flourish.

Contents

Foreword

Since 2017, techUK has played a pivotal role in convening discussions on digital ethics through our annual Digital Ethics Summit. Over the years, the landscape of digital ethics has evolved significantly, with the development of new ethical frameworks, establishment of institutions like The Ada Lovelace Institute, and DSIT's Responsible Technology Adoption Unit, and the growing community dedicated to addressing the complex challenges of digital ethics.

Recent developments in Artificial Intelligence (AI), including breakthroughs in Generative AI (genAI), global summits on AI governance, and emerging regulatory proposals, have brought ethical considerations at the forefront of technology policy. A key milestone in the UK was the government's 2024 AI White Paper[1], which introduced a principles-based, context-specific regulatory framework for AI, empowering the existing regulators to address AI-related risks within their respective sectors. The government now looks to build on this framework.

techUK has welcomed this approach, as it aligns with our calls for a pro-innovation, coordinated, and context-specific approach to AI regulation. In this context, closer regulatory cooperation, ensuring

that regulators are equipped with the necessary capabilities, and making sure the UK's approach is interoperable with international standards will be vital for ensuring the UK remains competitive and aligned with global AI developments.

We would be remiss not to acknowledge that the UK's approach exists within a broader global context – the European Union's AI Act is set to enter into force on 2 February 2025, while the United States has introduced The Executive Order for AI and the National Institute of Standards and Technology (NIST) AI Risk Management Framework, offering voluntary guidelines. China's strategy focuses on national security and technological independence, while the APAC region is developing diverse regulatory approaches. However, due to scope constraints, this paper will focus specifically on the UK's AI White Paper approach. Our aim is to provide in-depth insights into how the UK's principles-based framework can be effectively implemented by organisations operating within the UK regulatory landscape.

As the AI landscape continues to transform, organisations face a number of practical challenges in implementing AI ethical principles in real-world contexts. Balancing innovation with ethical considerations, interpreting broad principles in specific contexts, addressing the shortage of qualified personnel to implement and interpret tools, ensuring consistent application across diverse projects, and keeping pace with rapidly evolving AI technologies are just a few difficulties they must navigate. These challenges underscore the need for a practical, actionable framework to guide businesses through ethical AI implementation.

This paper recognises that regulatory guidelines are at times perceived as abstract or theoretical, while industry practices are seen as more concrete and applied. The aim of this paper is to bridge this gap, demonstrating how the regulatory principles, set out in the AI White Paper, are not merely theoretical constructs, but are already being operationalised within the industry. Through insights, illustrative tools, and real-world examples, the paper aims to offer information that organisations may find helpful as they consider how to apply the principles outlined in the AI White Paper in practice.

This paper is not a static document but a living resource open to feedback at this year's Digital Ethics Summit and contributions from the community, with plans to review and update it in 2025.

We hope it serves as a valuable insight for businesses, policymakers, and practitioners alike as they navigate this rapidly evolving landscape.

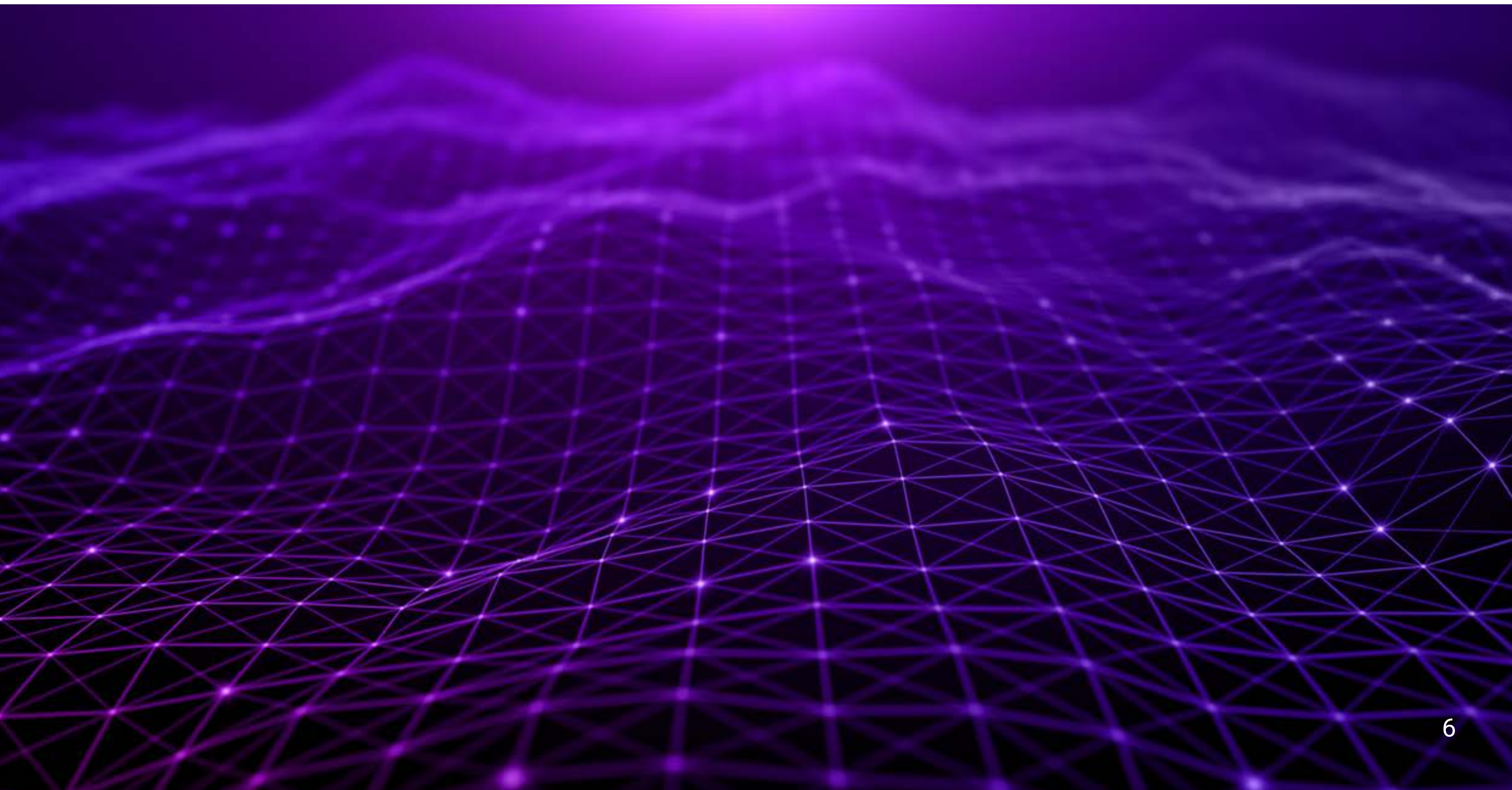
Sue Daley, Director,
Technology and Innovation
techK

Introduction

In the rapidly evolving landscape of AI regulation, the UK Government has taken a proactive approach to AI regulation through establishing the AI White Paper approach. The five ethical principles underpinning this White Paper – safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress – serve as a clear framework for what responsible AI should achieve.

The White Paper introduced a principles-based, agile, and context-specific approach to AI regulation, aimed at promoting responsible innovation. This framework delegates responsibility for AI regulation to existing regulators to address the risks and challenges within their respective sectors.

This paper explores how these principles can be, and in some cases are, being put into practice through the application of AI Assurance Techniques and Standards, often referred to as tools for trustworthy AI. It offers practical insights and examples from industry best practices and real-world scenarios that demonstrate successful achievement of each principle, illustrating how organisations might approach implementing these principles.



The regulatory principles outline what outcomes AI systems need to fulfil, while AI assurance techniques and standards offer concrete methods for demonstrating how a system has achieved these results in practice.

The paper is structured around the five principles of the UK AI White Paper. For each principle, we provide:

- An explanation of the principle and its implications;
- An overview of relevant tools and strategies for implementation, including specific AI assurance techniques and standards;
- Real-world case studies demonstrating how companies are currently applying these principles in practice.

The paper concludes with a summary table mapping various tools to each principle, offering a quick reference guide for organisations seeking to implement ethical AI practices.

The Five Ethical Principles Underpinning the UK's AI White Paper

The White Paper outlines five key ethical principles for responsible AI development and deployment. These principles are;

1. Safety, security and robustness
2. Appropriate transparency and explainability
3. Fairness
4. Accountability and governance
5. Contestability and redress

Together, they aim to ensure AI systems are not only technologically advanced but also align with societal values and ethical standards. Ethical AI practices build trust, mitigate risks, ensure legal compliance, offer competitive advantages, and contribute to societal benefits.

However, it is recognised that organisations face several obstacles in this journey of bridging the gap between ethical principles and their practical implementation. These include

difficulties in interpreting broad principles into specific operational guidelines, resource constraints (particularly for SMEs), the rapid pace of technological change, balancing innovation with ethical considerations, navigating the ever-evolving policy, standards and guidance landscape, and challenges in measuring adherence to ethical principles.

The tech industry itself has made significant strides in addressing these challenges. Many companies have developed innovative approaches to operationalise ethical principles, creating internal governance structures, training programs, and assessment tools. However, there's still work to be done to make ethical AI implementation more accessible and achievable for organisations of all sizes across all sectors. This paper aims to offer information that organisations may find helpful as they consider how to apply the principles outlined in the AI White Paper in practice to make ethical AI implementation more accessible and approachable.

Before exploring each of the AI White Paper Principles and how they can be implemented, it is important to first explain the tools we can use to achieve these principles, namely AI Assurance and Standards. The following section will explain what AI assurance and standards are and how they are key to operationalising ethical principles and ensuring their implementation.

Explaining AI Assurance and Standards as tools for trustworthy AI

AI Assurance Techniques and Standards serve as operational tools for translating ethical principles into tangible outcomes. These tools provide structured approaches for evaluating, verifying, and validating AI systems against the ethical principles outlined in the UK AI White Paper. They offer key benefits including objectivity, consistency, accountability, continuous improvement, and scalability in ethical AI implementation.

AI Assurance can be understood as mechanisms that support the verification, validation, and ongoing monitoring of AI development and deployment. It encompasses risk management frameworks, conformity assessments, audit and certification processes, and continuous monitoring techniques. These mechanisms allow companies to demonstrate compliance with ethical principles and regulatory requirements, build internal confidence in AI systems, and ensure the safety and reliability of AI systems.

Standards play a vital role in bridging the gap between high-level ethical principles and practical implementation. They provide specific, actionable guidelines for organisations to align their AI systems with ethical and regulatory requirements. The UK has been at the forefront of developing AI standards, contributing to both national and international efforts. This work includes BSI's development of AI-specific standards, contribution to ISO/IEC standards on AI, and sector-specific standards developed in collaboration with regulatory bodies.

These assurance techniques and standards offer a common language and set of best practices across the industry, serving as valuable tools for organisations seeking to implement ethical AI practices. They provide a framework for translating abstract principles into concrete, evidenced and measurable actions.

Having explored the White Paper's historical context, ethical principles, and the role of AI assurance and standards, we now turn to the paper's structure and purpose.

This paper examines each principle from the UK AI White Paper, providing an explanation of each principle, an overview of existing tools supporting their implementation, and presenting illustrative case studies of organizations currently applying these principles through AI assurance techniques and standards.

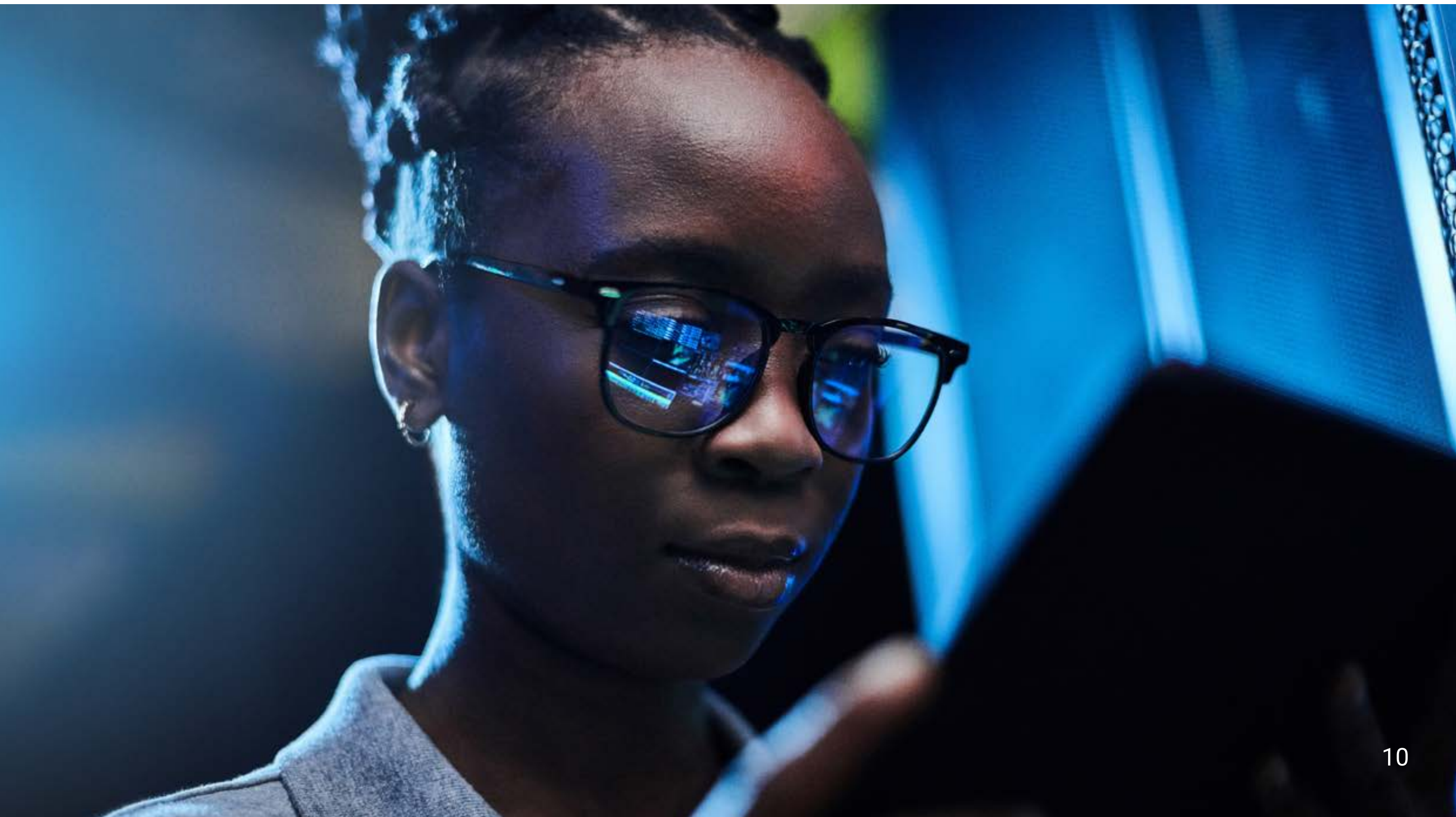
As we explore each principle, we encourage readers to consider how these guidelines could be applied within their own contexts. Our goal is to support organisations at all stages of their AI ethics journey to translate the regulatory principles set out in the AI White Paper into practical, actionable steps. By showcasing real-world examples, tools, and industry insights, we aim to illustrate how these principles are already being operationalised within the industry. This paper encourages organisations to explore available assurance and standards to support their AI ethics implementation.

Principle 1: Safety, Security and Robustness

The first principle of the UK AI White Paper, “Safety, security and robustness,” emphasises that AI applications should function in a secure, safe, and robust manner with careful risk management. This principle is fundamental to building trust in AI systems and ensuring their responsible deployment across various sectors.

Safety in AI refers to the system’s ability to operate without causing harm to users. This encompasses both physical safety in AI-controlled physical systems and broader societal safety considerations. Security focuses on protecting AI systems from unauthorised access, data breaches, and malicious attacks that could compromise their integrity or the privacy of user data. Robustness relates to an AI system’s ability to maintain consistent performance and reliability under varying conditions, including when faced with unexpected inputs or environmental changes.

The integration of these three aspects - safety, security, and robustness - is crucial for



developing AI systems that can be trusted and deployed responsibly. This principle underscores the need for comprehensive risk assessment and management throughout the AI lifecycle, from design and development to deployment and ongoing operation.

Implementing this principle requires a multifaceted approach that includes rigorous testing, continuous monitoring, and adaptive risk management strategies. It also necessitates consideration of potential long-term and indirect impacts of AI systems, beyond their immediate operational scope.

Tools for Implementing Safety, Security and Robustness

To support the implementation of this principle, organisations can draw on various tools and frameworks, for example:

Model verification techniques, which involve rigorous testing of AI models to ensure they perform as intended across a wide range of scenarios. Techniques such as **formal verification** can be used to mathematically prove that a model meets certain specifications.

Adversarial testing, which involves deliberately attempting to “fool” or “break” AI systems to help identify vulnerabilities and improve the system’s robustness.

Data modeling techniques help in understanding data distributions, identifying potential biases, data drift measurements and maintaining data integrity, which is crucial for the quality, representativeness, and security of training data.

The following are relevant standards that can support implementation of this principle:

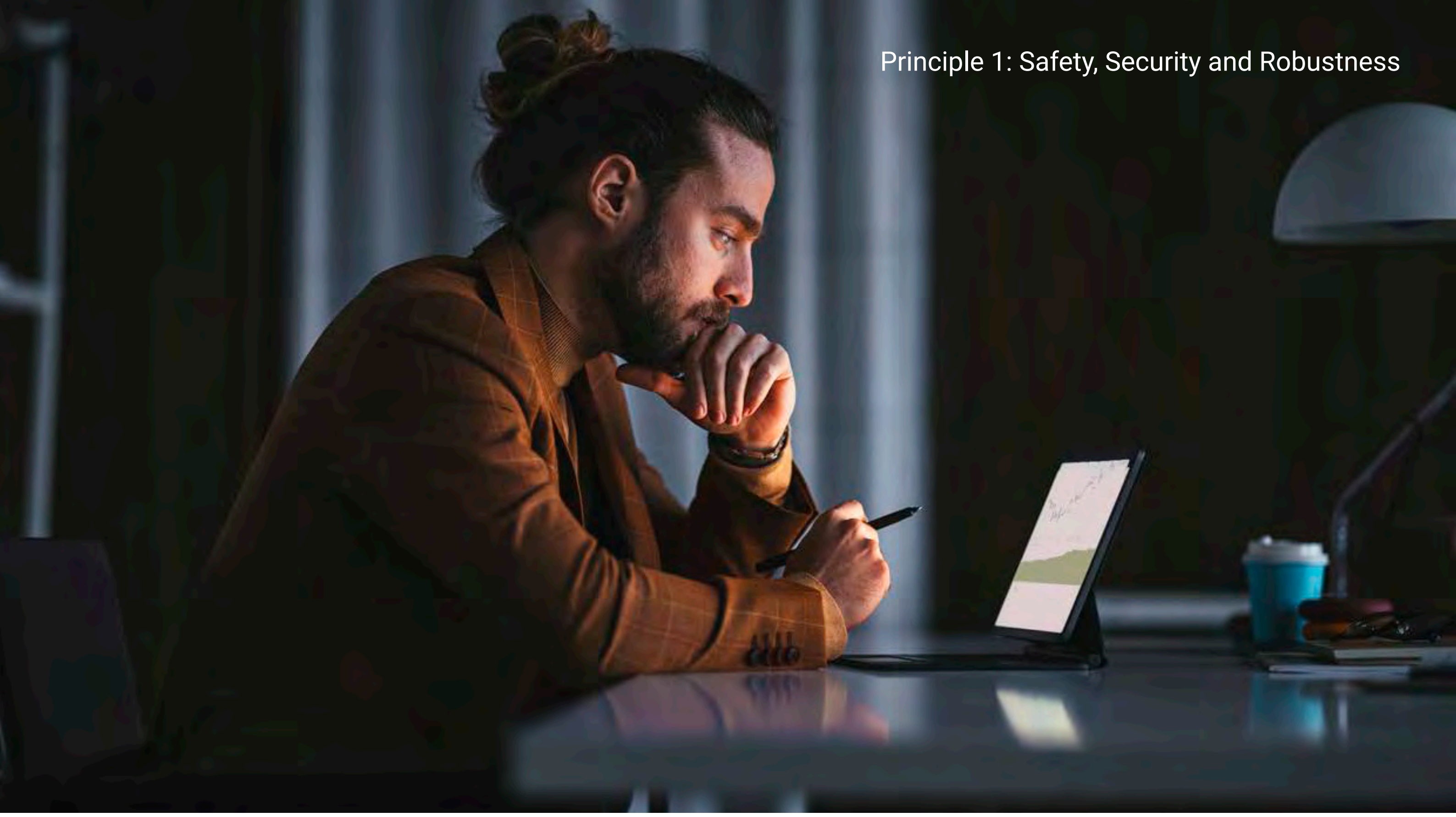
ISO/IEC 24029-2[2] AI — Assessment of the robustness of neural networks on AI trustworthiness;

ISO/IEC 27001:2022[3] - Information security, cybersecurity and privacy protection — Information security management systems, provides requirements for establishing, implementing, maintaining, and continually improving an Information Security Management System.

ISO/IEC 5259 AI — Data quality for analytics and machine learning (ML) Part 1[4]: Overview, terminology, and examples; Part 3[5]: Data quality management requirements and guidelines on data quality for analytics and machine learning; Part 4[6]: Data quality process framework which provides guidelines for improving data quality.

ISO/IEC TR 5469[7] AI — Functional safety and AI systems - functional safety and AI systems provide valuable guidance for implementing safety, security, testing, data quality, and robustness in AI systems.

Another important standards initiative to consider include CEN-CENELEC, a European standardisation body that develops and maintains voluntary technical standards across various industries to promote interoperability, safety, and innovation.



Given their crucial role in the standards that will support the EU AI Act and their success in transforming European standards into global ones; and United States's NIST (National Institute of Standards and Technology), which is particularly significant for AI organisations in the United States. By leveraging these international initiatives, organisations can support their risk management practices, ensuring alignment with global standards and improving resilience against emerging risks.

Organisations might consider implementing comprehensive **risk management frameworks** such as the National Institute of Standards and Technology (NIST) Risk Management Framework[8] that cover the entire AI lifecycle to help maintain safety, security, and robustness.

Additionally, conducting **impact assessments** is a systematic evaluation of the potential effects and risks associated with implementing AI systems in an organisation or society. This tool can prove valuable for identifying potential benefits and harms, helping decision-makers make informed choices about AI deployment and governance.

Example of Industry Use of Assurance and Standards to Achieve the Principle of Safety, Security and Robustness

Case Study: Kainos, Conducting Ethics and Harm Workshops for Defence AI Projects

Kainos, engaged by the Defence Science and Technology Laboratory (Dstl), served as the delivery partner for the Defence AI Centre (DAIC) programme of advanced rapid AI experimentation. When designing for military scenarios where ethics and trust are paramount, the approach went beyond exploring potential harms of AI-enabled systems in their intended use.

A key component of Kainos' methodology within the DAIC was conducting ethics and harm workshops, which informed both impact assessments and design choices. Guided by a data ethicist, teams defined potential benefits, harms, and mitigation strategies for AI-enabled systems or services. These workshops were structured around the MoD AI ethical principles: human-centricity, responsibility, understanding, bias and harm mitigation, and reliability.

Safety, security, and robustness challenges were explicitly addressed through the principles of reliability and human-centricity. The workshops explored scenarios such as the system being used in a different context to which it was originally designed, risks of unintentional misuse or abuse, and potential consequences of deployment in strategically sensitive environments. These insights typically informed testing strategies for the AI system under development.

By integrating these workshops at the start of the agile delivery cycle and revisiting them at subsequent stages, an ethics-by-design approach could be ensured. This process allowed for continuous updating and supplementation of mitigation strategies as project parameters evolved. The ethics and harm workshops formed an integral part of a comprehensive delivery framework, guiding structural checkpoints, safety and legal considerations, and testing protocols.

Principle 2: Appropriate Transparency and Explainability

The second principle of the UK AI White Paper, *"Appropriate transparency and explainability,"* emphasises that organisations developing and deploying AI should communicate when and how AI is used and provide explanations of a system's decision-making process at a level of detail that corresponds to the risks posed by the AI's use.

This principle is crucial for building trust in AI systems and ensuring accountability. Transparency in AI refers to the openness about when and how AI is used, enabling stakeholders to understand the extent of AI's influence in various processes. Explainability, on the other hand, focuses on providing insights into how AI system arrives at its outputs or decisions, making their inner workings more understandable to users and regulators.

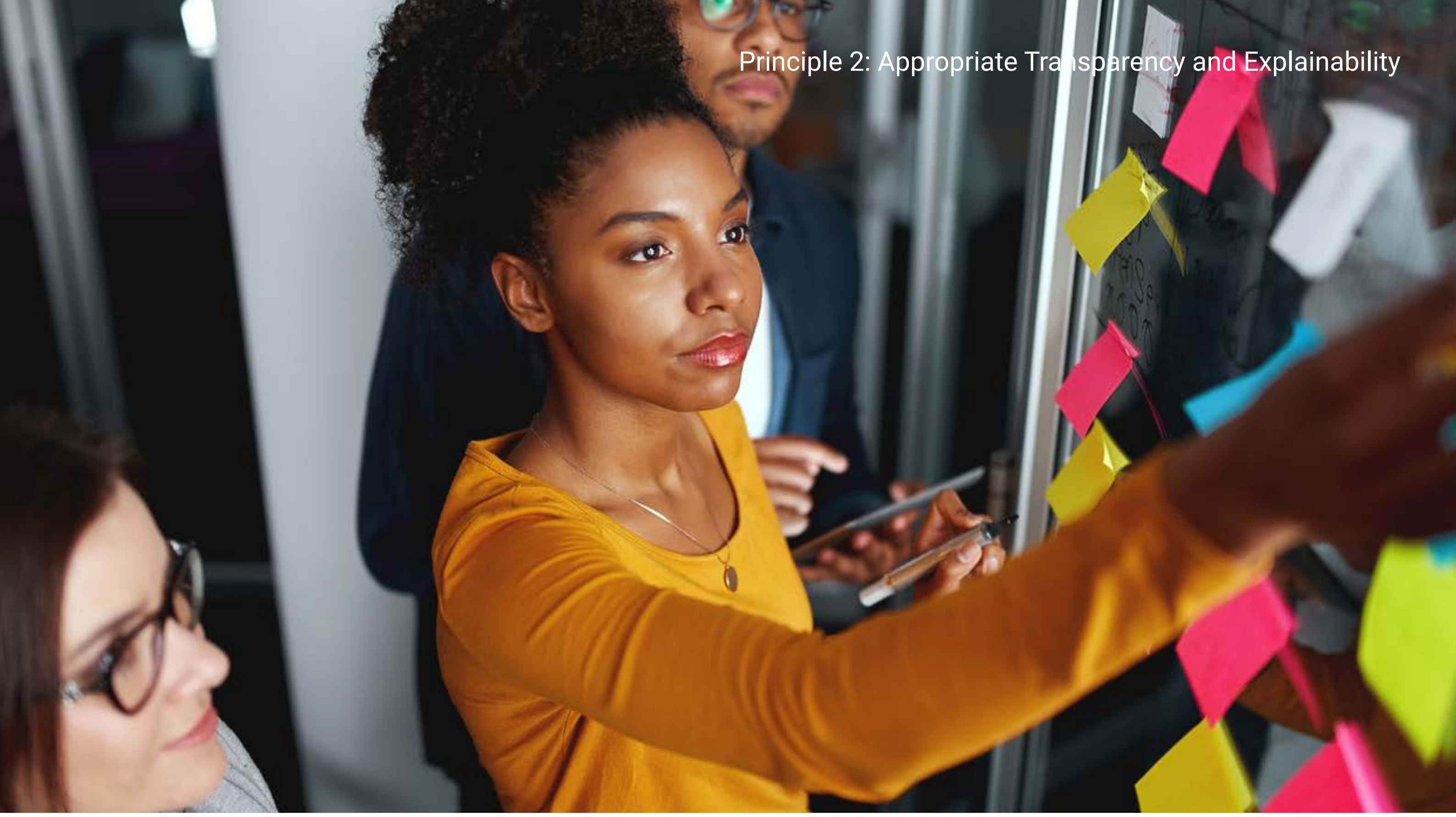
The concept of "appropriate" is key here, recognising that the level of transparency and explainability should be commensurate with the potential risks and impacts of the AI system. For low-risk applications, a basic understanding of the AI's role might suffice. However, for high-risk applications, such as those in healthcare or criminal justice, regulatory frameworks may call for a more detailed explanation of the AI's decision-making process to ensure accountability and transparency.

Implementing this principle helps address concerns about AI bias, fairness, and accountability. It enables stakeholders to understand and, when necessary, challenge AI-driven decisions, fostering responsible and ethical use of AI technology. Moreover, appropriate transparency and explainability can support regulatory compliance and continuous improvement of AI systems.

Achieving explainability in complex AI systems, particularly deep learning models, poses significant challenges. It requires balancing transparency with system performance and intellectual property concerns. This balance often requires careful consideration and innovative approaches.

Transparency is further complicated by AI supply chain dynamics. Decisions made by developers regarding model architecture, training data, and algorithmic design can influence outcomes in ways that may not always be fully visible to deployers or end users. At the same time, developers may have limited visibility or control over how their systems are ultimately applied, potentially making it difficult to anticipate all potential downstream uses, or mitigate risks arising from unintended or harmful applications, including malicious misuse.

Black box AI refers to AI systems whose internal workings and decision-making processes are not transparent or easily interpretable by humans. Black box AI approaches, while offering powerful capabilities, present challenges in terms of trust, debugging, and potential biases. In contexts involving critical decisions or regulatory oversight, the lack of interpretability may raise ethical and practical concerns. Striking a balance between leveraging the advantages of complex AI models and ensuring sufficient transparency remains a key challenge in developing responsible AI systems.



These considerations necessitate a holistic view of AI development and deployment to achieve meaningful explainability and accountability throughout the entire lifecycle and usage of AI systems.

Tools for Implementing Appropriate Transparency and Explainability

Many organisations have established procedures for notifying data subjects about the processing of their personal data, as required by the UK GDPR. These processes can provide a useful foundation that organisations may adapt and expand to cover the specific requirements of AI systems.

To further support transparency and explainability in AI, organisations can also employ a range of additional tools and strategies as and where appropriate. These include:

Setting clear expectations for those involved in developing and deploying AI systems to proactively or retrospectively provide information about the AI system. This includes details about **the nature and purpose of the AI,** including information relating to any specific outcome. It also encompasses transparency about the data being used, including information relating to training data.

Organisations can aim to provide clear explanations of the logic and processes used by their AI systems, providing information to support the explainability of decision-making and outcomes where relevant. This might involve the **use of interpretable AI models** or the application of post-hoc explanation techniques for more complex models. Additionally, **clear accountability** for the AI and any specific outcomes can be established and communicated.

For higher-risk systems, more stringent explainability requirements can be set. This ensures an appropriate balance between the information needs for regulatory enforcement (for example, around safety) and technical trade-offs with system robustness. Tools like **model cards**, which provide structured information about an AI model's characteristics and limitations, can be useful in this context.

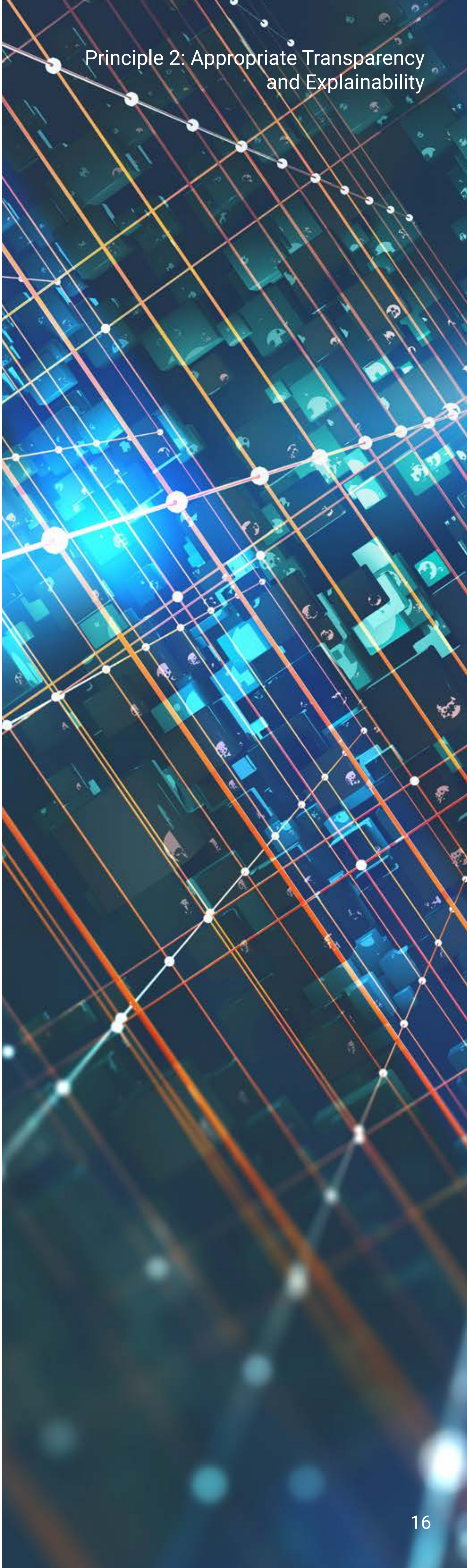
Similarly, **system cards** offer a broader view, documenting not just the model but the entire AI system, including its intended use, performance metrics, and potential societal impacts. These tools help standardise and communicate critical information about AI systems to various stakeholders.

Several **technical standards** addressing AI transparency and explainability can be leveraged to support implementation. These include

IEEE 7001[9] Standard for Transparency of Autonomous Systems, which provides a framework for transparency in autonomous systems,

These standards can help clarify regulatory guidance and support the implementation of risk treatment measures.

Also, techniques such as **SHAP (SHapley Additive exPlanations**) values, **LIME (Local Interpretable Model-agnostic Explanations)**, and attention visualisation for neural networks can provide insights into model decision-making. For simpler models, decision trees or rule-based systems can offer inherent explainability.



Example of Industry Use of Assurance and Standards to Achieve the Principle of Appropriate Transparency and Explainability

Case Study: GreenhouseAI, Implementing "License to Operate" for AI Model Transparency in Financial Services

GreenhouseAI supports a UK-based financial services provider, which holds data on over 30 million UK adults, in ensuring transparency and ethical compliance of their AI models. The organisation has implemented robust governance and ethical frameworks to manage their sensitive and personal data effectively.

A key element of these frameworks is the "License to Operate" (LtO), based on established principles of model cards. The LtO provides a standardised method for documenting crucial model information, training data, and performance metrics, thereby ensuring transparency, accountability, and fairness in AI development and use.

Each model with a distinct intended use receives its own LtO, even if it's a derivative of a previous model. The LtO compilation begins during the development process once a model is deemed 'viable' for production. Much of the information gathering has been automated using MLOps tools. The LtO includes sections on model details, intended use, performance metrics, training data, production data, ethical considerations, and caveats.

The LtO remains with the model throughout its lifecycle, updated as necessary, such as during retraining. The key principle behind the LtO is to ensure that anyone can understand everything about the model, especially its usage and ethical limitations, even if all original team members have left the company.

This tool has become a crucial and mandatory part of the organisation's AI governance processes, supporting GreenhouseAI's commitment to ethical and transparent AI implementation in the financial services sector.

Principle 3: Fairness

The third principle of the UK AI White Paper, "Fairness," emphasises that AI should be used in a way that complies with existing UK laws, such as the Equality Act 2010 and UK GDPR, and must not discriminate against individuals or create unfair market outcomes. This principle is fundamental to ensuring that AI systems do not perpetuate or exacerbate societal biases and inequalities.

Fairness in AI, as outlined in the AI White Paper, is a multifaceted concept that goes beyond mere compliance with anti-discrimination laws. It requires a nuanced understanding of how AI systems can impact different groups and individuals, both directly and indirectly. The White Paper calls for actors involved in all stages of the AI to "consider definitions of fairness that are appropriate to a system's use, outcomes and the application of relevant law[10]," recognising that the definition of fairness may vary depending on the context and application of the AI system.

Moreover, as presented in the AI White Paper, organisations must decide the contexts and instances where fairness is particularly relevant, acknowledging that it may not always be applicable or may need to be balanced against other considerations. This requires a thoughtful approach to AI development and deployment, with consideration of potential impacts on various stakeholders.

Promoting fairness in AI systems involves considering factors like data quality, potential algorithmic bias, and unintended consequences. It also requires considering the broader societal implications of AI systems, including their potential to reinforce existing inequalities or create new ones.

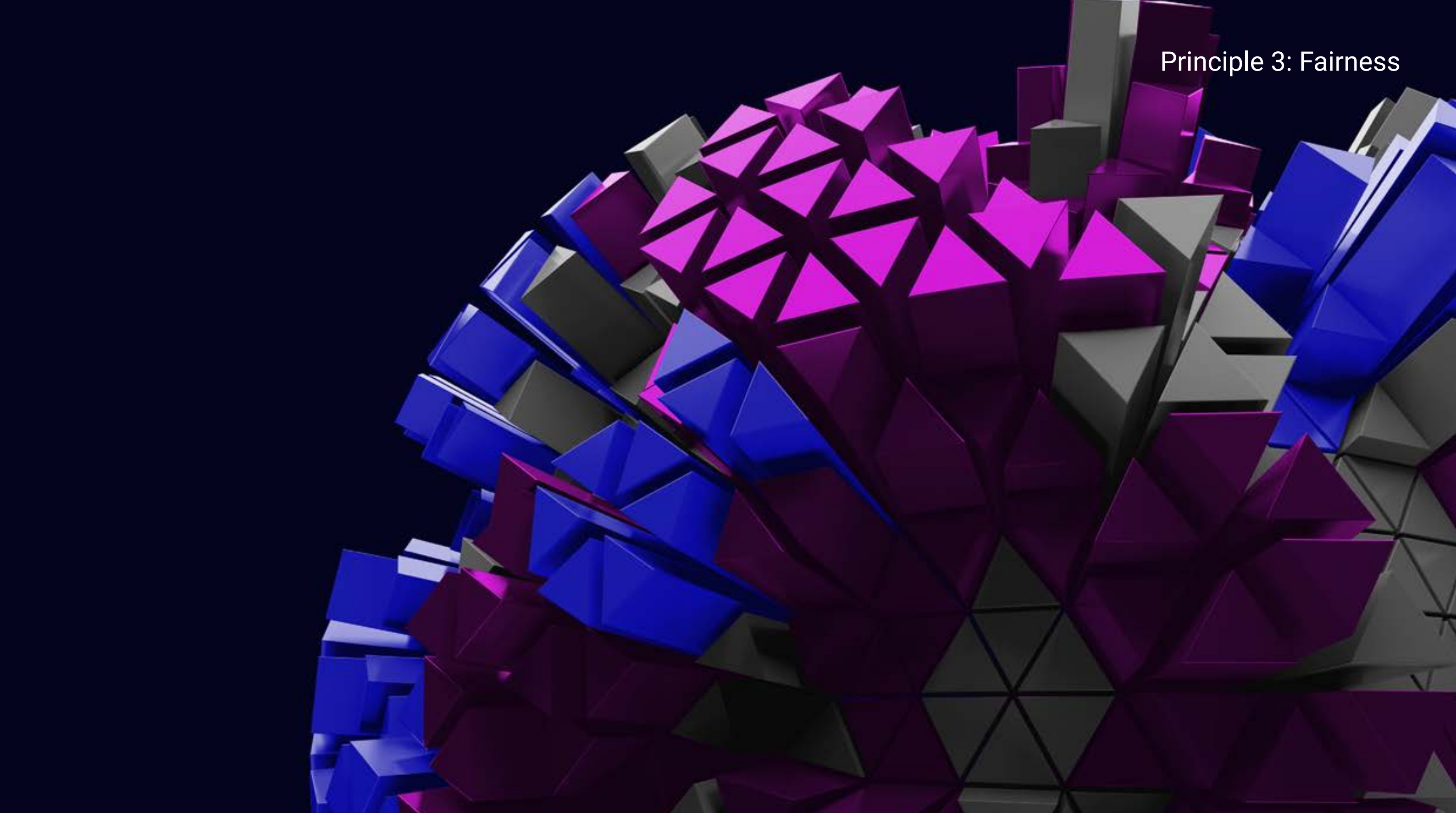
Fairness can be considered throughout the AI lifecycle, from data collection and model development to deployment and monitoring. Even before beginning development, developers can ask themselves, "Is this the right thing to do?" to assess the ethical implications and potential societal impact of their AI project. This critical reflection at the outset helps ensure that fairness and ethical considerations are fundamental to the AI's purpose and design. Similarly, those considering deploying AI solutions can first ask, "What problem am I trying to solve, and is AI the right approach?" This ensures that AI is used judiciously and when it's the most appropriate solution to the problem at hand.

Tools for Implementing Fairness

To implement the principle of Fairness, organisations are employing a range of various tools and strategies. The following provides more details on the tools, strategies and frameworks that can support this principle, these include:

Where a decision involving the use of an AI system has a legal or similarly significant effect on an individual, system operators might consider being prepared to **offer an appropriate explanation for that decision to affected parties.** This approach could promote transparency and accountability, potentially allowing individuals to better understand and, if necessary, question decisions that affect them.

It can be beneficial for AI systems to align with regulatory guidelines relating to the vulnerability of individuals within specific regulatory domains. This could involve **reflecting on how the use of AI systems might influence individuals' vulnerability,** in line with existing regulatory frameworks. For instance, in financial services, AI systems used for credit decisions



might take into account their potential impact on financially vulnerable individuals.

Several **technical standards** addressing AI fairness, bias mitigation, and ethical considerations can be leveraged to support implementation. These include the following:

ISO/IEC TR 24027:2021[11] Information technology — AI — Bias in AI systems and AI aided decision making, provides guidance on bias in AI systems and AI-aided decision making.

ISO/IEC TR 24368:2022[12] Information technology — AI — Overview of ethical and societal concerns.

These standards can help clarify regulatory guidance and support the implementation of risk treatment measures.

Organisations can also use **fairness metrics** to assess and monitor their AI systems. These might include demographic parity, equal opportunity, and equalised odds, there are over 20 different fairness metrics. However, it's important to note that different fairness metrics may be appropriate in different contexts, and they can sometimes be in tension with each other.

Bias detection and mitigation tools can be employed during the development and testing phases of AI systems. These tools can help identify potential biases in training data or model outputs, allowing developers to address these issues before deployment.

Here, it is worth touching upon the **historical biases**, which occur when past societal prejudices and inequalities are inadvertently embedded in modern data, leading to biased outcomes in AI decision-making processes. To address historical bias in AI, several effective solutions can be employed. Techniques like **reweighting and resampling** adjust training data to ensure underrepresented groups are given appropriate emphasis, helping models better reflect diverse populations. Bias detection tools can support the identification of disparities in how different groups are treated, providing insights that allow for targeted adjustments.

Federated learning, a machine learning approach where models are trained on decentralised data, can help reduce bias by allowing diverse datasets to be used without compromising privacy or requiring data centralisation. **Regular fairness audits** of AI systems can help ensure ongoing compliance with fairness principles. These audits can consider both the technical aspects of the AI system and its real-world impacts on different groups.

Example of Industry Use of Tools to Achieve the Principle of Fairness

Case study: Sopra Steria, Monitoring Demographic Diversity in MRI Trials

Sopra Steria collaborated with a medical device startup to develop, test, and engineer AI technology aimed at significantly reducing patient time in MRI machines for cancer detection. Recognising potential ethical concerns around data protection and cross-cultural applicability, the team engaged Sopra Steria's Ethics and Sustainability consulting team.

The consulting team worked closely with radiographers, radiologists, medical scientists, and data scientists to understand the problem and proposed solution. Their review identified several areas for consideration, including potential bias, privacy, fairness, and environmental concerns.

The ethics review revealed that the project had already addressed many typical ethical concerns, such as job loss and patient engagement. However, a critical insight emerged: given that MRIs image internal organs, the project team had not initially considered ethnicity bias as a potential issue. Although there was no evidence of gender or ethnicity bias in MRIs at the time, the project team agreed to monitor gender, age, and ethnicity of trial participants to proactively demonstrate the AI's unbiased performance.

This decision proved timely, as later that year, medical journals published findings that some datasets and resulting AIs developed for MRIs were showing both ethnic and gender bias. Sopra Steria's proactive approach to demographic monitoring in their trials positioned them to address these emerging concerns in the field of AI-assisted medical imaging.

By implementing this monitoring system, Sopra Steria demonstrated its commitment to developing ethical and unbiased AI solutions in healthcare, setting a standard for responsible innovation in medical technology.

Case Study: Epic, Developing Assurance Suite for Equitable Healthcare

Epic, an Electronic Patient Record vendor serving numerous NHS Trusts, has taken a proactive approach to ensuring AI-driven features support fair and equitable care for diverse populations. Recognising that one-time validation efforts were insufficient to prevent bias in probabilistic AI tools, Epic developed a comprehensive solution for ongoing performance monitoring. The company's AI Trust and Assurance Suite provides healthcare organisations with near real-time metrics and analysis of AI models' performance on their specific patient populations. This approach addresses the challenge of local validation, which has historically been labour-intensive and required resources not available to all healthcare providers.

The AI Trust and Assurance Suite features automated, intuitive reporting dashboards that eliminate the need for in-house data mapping. It offers analysis broken down by age, sex, race/ethnicity, and other demographics, ensuring continuous monitoring of AI-driven tools' performance over time. Because the suite can evaluate data specific to each organisation, it provides value across diverse healthcare settings, from rural hospitals to specialty pediatric facilities.

In a move towards greater transparency and collaboration, Epic has made the suite's monitoring template and data schema publicly available on GitHub. This allows healthcare organisations to monitor their own custom AI models, evaluate models purchased from third-party vendors, and contribute their own metrics to the broader healthcare community. By developing this suite, Epic demonstrates its commitment to equitable healthcare delivery and responsible AI implementation. The tool empowers healthcare providers to ensure their AI-driven features perform consistently across diverse communities. This initiative sets a new standard for transparency and accountability in healthcare AI, facilitating the responsible adoption of AI technologies that can improve patient outcomes and reduce clinician workload.

Principle 4: Accountability and Governance

The fourth principle of the UK AI White Paper, "Accountability and governance," emphasises the need for appropriate oversight of AI usage and clear accountability for outcomes. This principle is crucial for ensuring responsible development, deployment, and use of AI systems.

Accountability in AI refers to the obligation of an organisation and its members to report, explain, and be answerable for the consequences of their AI-driven decisions and actions. It involves taking responsibility for the impacts of AI systems, both intended and unintended. Governance, on the other hand, encompasses the structures, processes, and practices put in place to ensure proper development, deployment, and use of AI systems.

Effective accountability and governance in AI require a clear chain of responsibility within organisations. This includes defining roles and responsibilities for AI development, deployment, and monitoring, as well as establishing mechanisms for addressing issues or concerns that arise. It also involves creating transparent processes for decision-making around AI systems and their use.

Implementing this principle helps build trust in AI systems by ensuring that there are clear lines of responsibility and recourse in case of problems. It also promotes responsible innovation by encouraging organisations to carefully consider the potential impacts of their AI systems before deployment.

However, achieving robust accountability and governance in AI can be challenging due to the complex and often opaque nature of AI systems, especially in cases where decision-making processes are not easily interpretable. Organisations can consider the development of new approaches and adapt existing governance structures to effectively manage AI-related risks and responsibilities.

Tools for Implementing Accountability and Governance

There are numerous ways to support accountability and governance in the implementation of AI systems. While the maturity of assurance may vary depending on an organisation's sector or size, a wide range of tools and strategies are available for reference. The following section provides an overview of the tools and frameworks that support this principle, including:

Organisations have the opportunity to establish **clear governance mechanisms**, including activities within the scope of appropriate **risk management and governance processes**. This may include reporting duties to ensure transparency and oversight. Governance structures can be designed to cover the entire

AI lifecycle, from development and testing to deployment and ongoing monitoring. Effective AI governance frameworks can strike a balance between enabling innovation and maintaining appropriate controls, encompassing the full AI lifecycle while allowing for the iterative nature of AI development within a structured risk management approach.

Several **technical standards** addressing AI governance, risk management, and transparency can support responsible behavior and maintain accountability within an organisation. These include the following:

ISO/IEC 23894[13] Information technology — AI — Guidance on risk management provides guidelines for AI risk management, offering a framework for identifying, assessing, and mitigating AI-related risks.

ISO/IEC 42001[14] Information technology — AI — Management system focuses on AI management systems, providing a structured approach to AI governance.

ISO/IEC 5469[15] AI — Functional safety and AI systems addresses functional safety and AI systems

ISO/IEC 25059[16] Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems provides guidance on systems and software quality requirements and evaluation for AI systems.

Organisations can also implement **AI ethics boards or committees** to provide oversight and guidance on AI development and deployment. These bodies can review proposed AI projects, assess their potential impacts, and ensure

alignment with ethical principles and regulatory requirements. To support transparency and accountability, organisations can develop and publicly disclose their AI ethics principles, modeled after established frameworks like the UK AI White Paper, while also establishing AI ethics boards to provide ongoing oversight and ensure adherence to these principles throughout the AI lifecycle.

Regular **AI audits and impact assessments** can help organisations maintain accountability by systematically evaluating the performance and broader impacts of their AI systems. These assessments should consider not only technical performance but also broader societal and ethical implications. Deployers of existing AI models may have limited visibility into the development and training processes of these models, as well as their inner workings. This can pose challenges in obtaining detailed information about the model's training data, potential biases, and limitations, making it difficult to maintain comprehensive and fully transparent documentation. Nonetheless, implementing robust documentation practices is also crucial for accountability. This includes maintaining detailed records of AI system development, training data, decision-making processes, and any issues or incidents that arise which could be captured in a model card.

Examples of Industry Use of Assurance and Standards to Achieve the Principle of Accountability and Governance

Case Study: Trilateral Research, Implementing an Accountability Framework for AI in Child Safeguarding

Trilateral Research has established a comprehensive accountability framework for its AI-enabled system, CESIUM®, which enhances decision-making in safeguarding children at risk of criminal and sexual exploitation. This framework ensures responsibility and accountability throughout the tool’s development and use.

The company maintains detailed records of design, development processes, and decision-making. A multidisciplinary sociotech team, comprising ethics experts, subject matter specialists, and technical scientists, regularly evaluates CESIUM’s ethical impact. With continuous leadership support, Trilateral Research implements an ethics-by-design approach to produce an ethical product.

As part of its licensing contract, Trilateral Research has established a shared responsibility framework with clients. The company takes responsibility for establishing transparency in data collection, processing, and scope of use; identifying and mitigating data bias; providing clear lines of accountability; protecting the privacy of human subjects; and encouraging end users’ understanding and professional assessment of the product’s output.

Recognising that AI tools can be used in ways that either promote or undermine ethical values and fundamental rights, Trilateral Research also outlines client responsibilities. These include integrating the product into their organisation, complying with applicable laws and legislation, investigating potential operational biases, continuously assessing AI output, understanding the nuances of the operational environment, and explaining to affected stakeholders how the tool informed decision-making.

To support this shared accountability, Trilateral Research maintains an open line of communication with clients, discussing their responsibilities, ethics roadmaps, bias assessments, algorithmic audits, and ethics awareness training. This comprehensive approach ensures that CESIUM® is developed and used responsibly, prioritising ethical considerations in the sensitive area of child safeguarding.

Case Study: MeVitae, Implementing ISO 27001 in HR Technology

MeVitae, an organisation specialising in HR technology, has integrated AI assurance principles into its core technology and processes to enhance transparency, fairness, and accountability in recruitment.

A significant milestone in MeVitae's journey was its participation in an ICO audit, where the company proactively invited assessment of its AI practices. This initiative led to the ICO developing an AI audit process modeled on MeVitae's work, highlighting the company's commitment to security (demonstrated through ISO 27001 certification) and its established AI ethics process.

MeVitae employs various assurance mechanisms to ensure the trustworthiness of its recruitment AI systems, including bias audits, compliance audits, and risk assessments. These processes align with the RTA's guidance on the lifecycle of AI assurance. The company collects quantitative data on its AI-driven systems' performance, ensuring fairness and transparency in hiring decisions. MeVitae regularly assesses the impact of its AI systems against industry standards and benchmarks to address biases and ensure system robustness. Furthermore, the company transparently shares its assurance processes and evaluation results with both internal and external stakeholders, demonstrating alignment with ethical and regulatory principles.

MeVitae's proactive approach, including its ICO audit participation and ISO 27001 certification, showcases the company's leadership in navigating complex regulatory landscapes like GDPR and the Equality Act 2010. By combining AI ethics with practical assurance tools, MeVitae sets a standard for trustworthy recruitment technology while adapting to evolving regulatory environments, helping organisations build inclusive and transparent hiring processes.

Principle 5: Contestability and Redress

The fifth, and final, principle of the UK AI White Paper, “Contestability and redress,” emphasises that people need to have clear routes to dispute harmful outcomes or decisions generated by AI. This principle is crucial for maintaining public trust in AI systems and ensuring that individuals have recourse when AI-driven decisions negatively impact them.

Contestability refers to the ability to challenge or question the outputs or decisions made by an AI system. This is particularly important in contexts where AI systems make or influence significant decisions affecting individuals’ lives.

Redress relates to the mechanisms in place for individuals to seek remedies or corrections when they have been unfairly treated or harmed by an AI system’s decision. This could involve correcting erroneous data, revising a decision, or compensating for harm caused.

Implementing the principle of contestability and redress requires organisations to create transparent processes for individuals to understand, question, and if necessary, challenge AI-driven decisions. It also necessitates the establishment of clear pathways for addressing grievances and providing appropriate remedies.

This principle is closely linked to other ethical principles, particularly transparency and explainability. Without sufficient understanding of how an AI system operates and makes decisions, it becomes difficult for individuals to effectively contest its outputs or seek meaningful redress.

This principle also plays a crucial role in maintaining human agency in AI-driven processes. By ensuring that AI decisions are not final or unchallengeable, it preserves the important role of human judgment and oversight in significant decision-making processes.

Tools for Implementing Contestability and Redress

To implement the principle of contestability and redress, organisations can employ various tools and strategies.

A crucial first step is creating or updating guidance with relevant information on where to direct a complaint or dispute for those affected by AI harm. This guidance can support clarity in existing **‘formal’ routes of redress** offered in certain scenarios, ensuring that individuals know where to turn when they need to contest an AI-driven decision.

Organisations can also clarify the interactions between contestability and redress and the requirements of appropriate transparency and explainability. These latter principles act as pre-conditions for effective redress and contestability. Without clear explanations of how AI systems reach their decisions, it becomes challenging for individuals to meaningfully contest those decisions.

Implementing **'human-in-the-loop' systems** can be an effective tool for contestability. These systems ensure that significant AI-driven decisions are reviewed by human experts before being finalised, providing an initial point of contestability within the decision-making process itself.

Organisations can establish dedicated **AI complaint handling processes**, separate from general customer service channels. These specialised processes can be equipped to handle the unique challenges of AI-related complaints, including the need for technical explanations and potential biases.

Regular audits of AI systems, focusing not just on performance but also on the fairness and contestability of outcomes, can help organisations proactively identify and address potential issues before they lead to complaints.

Collaboration with external stakeholders, including consumer rights groups, independent advisors and regulatory bodies, can help organisations develop more robust and effective contestability and redress mechanisms. These collaborations can provide valuable insights into the types of issues that are likely to arise and the most effective ways to address them.

Organisations can also consider implementing **'algorithmic recourse'** - the ability for individuals to understand what they would need to change to receive a different decision from the AI system. This can be particularly valuable in contexts like loan applications, school grades or job recruitment.

Example of Industry Use of Assurance and Standards to Achieve the Principle of Contestability and Redress

Case Study: VE3 Implementing AI Decision Review Portal and Bias Audits for Financial Services

VE3 partnered with a global financial services provider to implement robust contestability and redress mechanisms within its AI systems, ensuring ethical oversight and accountability throughout the AI lifecycle. The company’s approach emphasises human review of AI-driven decisions before finalisation, particularly for critical choices such as resource allocation and loan approvals. This process helps mitigate potential biases and unfair outcomes by subjecting AI decisions to thorough scrutiny.

Central to VE3’s strategy is the AI Decision Review Portal, a formal process established for employees and stakeholders to contest AI decisions. This portal played a crucial role when the AI system inaccurately classified several applicants as high-risk. Affected individuals were able to utilise the portal to submit additional context, prompting a review of the AI’s decision-making process. This mechanism ensures that AI decisions are not only contestable but also subject to human oversight and correction when necessary.

VE3 conducts regular bias audits of AI models to evaluate fairness and accuracy. These audits involve collaboration between internal teams and external stakeholders, ensuring transparency and continuous improvement in AI governance. By creating continuous feedback loops to enhance decision-making processes, VE3 fosters a culture of trust among employees and customers, reinforcing the organisation’s commitment to ethical AI practices.

The implementation of these contestability and redress mechanisms has significantly improved the financial services provider’s AI governance. It has enhanced transparency, accountability, and fairness in AI-driven decisions, particularly in sensitive areas like loan approvals and resource allocation. VE3’s approach demonstrates how financial institutions can leverage AI technology while maintaining strong ethical standards and building trust with their stakeholders.

Summary table: tools for trustworthy AI applied to principles

The following summary table serves as a valuable resource for organisations seeking to understand how the various tools and standards explored in this paper align with each principle, enabling companies to make informed decisions about their AI governance strategies.

The table is particularly beneficial for resource-constrained organisations, as it allows them to identify assurance mechanisms or standards that address multiple principles simultaneously, maximising the impact of their limited resources.

By visualising the cross-principle applicability of different tools, companies can prioritise their efforts and invest in solutions that offer the broadest coverage of ethical considerations. It's worth noting however that while a tool might be primarily associated with one or two principles, it could still have indirect benefits for others. The "Yes" indicates a strong, direct application to the principle, while "No" doesn't necessarily mean the tool is irrelevant, just that it's not a primary tool for that specific principle.

At the same time, it is important to note that some assurance tools, although they may only align closely with one principle, are of great importance. Further, it is worth noticing how all but one of the standards align specifically to one principle, showcasing their specificity.

This comprehensive view also helps organisations identify potential gaps in their current AI governance frameworks and guides them towards a more holistic approach to responsible AI development and deployment. Ultimately, this table aims to support companies in efficiently navigating the complex landscape of AI ethics and governance, facilitating more effective and principled AI implementations across various sectors.

The Y-axis is an alphabetical list of tools for trustworthy AI listed in the paper and the X-axis is the five principles which underpin the UK's White Paper.

	Principle 1: Safety, security and robustness	**Principle 2: Appropriate transparency and explainability**	**Principle 3: Fairness**	**Principle 4: Accountability and governance**	**Principle 5: Contestability and redress**
Assurance Mechanisms					
Adversarial testing	✓	✗	✗	✗	✗
AI Ethics Boards	✓	✓	✓	✓	✓
AI Impact Assessments	✓	✓	✓	✓	✗
Algorithmic Audits	✓	✓	✓	✓	✗
Bias Detection and Mitigation Tools	✗	✗	✓	✗	✗
Data Modelling	✓	✓	✓	✗	✗

	Principle 1: Safety, security and robustness	**Principle 2: Appropriate transparency and explainability**	**Principle 3: Fairness**	**Principle 4: Accountability and governance**	**Principle 5: Contestability and redress**
Assurance Mechanisms					
Explainable AI Techniques	✗	✓	✗	✗	✓
Fairness Metrics	✗	✓	✓	✗	✗
Human-in-the-loop Systems	✓	✗	✓	✗	✓
Model Cards	✗	✓	✓	✓	✓
Model Verification	✓	✗	✗	✗	✗
Risk Management Frameworks	✓	✗	✗	✓	✗
System Cards	✓	✓	✓	✓	✓

	Principle 1: Safety, security and robustness	**Principle 2: Appropriate transparency and explainability**	**Principle 3: Fairness**	**Principle 4: Accountability and governance**	**Principle 5: Contestability and redress**
Standards					
ISO/IEC 24029 Assessment of the robustness of neural networks on AI trustworthiness	✓	✗	✗	✗	✗
ISO/IEC 5259 Data quality for analytics and machine learning	✓	✗	✗	✗	✗
ISO/IEC TR 5469 Functional safety and AI systems	✓	✗	✗	✗	✗
IEEE 7001 Standard for Transparency of Autonomous Systems	✗	✓	✗	✗	✗
ISO/IEC TR 24027:2021 Bias in AI systems and AI aided decision making	✗	✗	✓	✗	✗
ISO/IEC TR 24368:2022 Overview of ethical and societal concerns	✗	✗	✓	✗	✗
ISO/IEC 23894 Guidance on risk management provides guidelines for AI risk management	✗	✗	✗	✓	✗

	Principle 1: Safety, security and robustness	**Principle 2: Appropriate transparency and explainability**	**Principle 3: Fairness**	**Principle 4: Accountability and governance**	**Principle 5: Contestability and redress**
Standards					
ISO/IEC 42001 Management system focuses on AI management systems	✓	✓	✓	✓	✓
ISO/IEC 25059 Systems and software Quality Requirements and Evaluation (SQuaRE)	✗	✗	✗	✓	✗
ISO/IEC 27001 Information security management systems	✓	✗	✗	✓	✗

To add an assurance mechanism or standard to this RAG framework that you are using to achieve one of the UKs White Paper Principles or to contest one of the yes/no labels, please email tess.buckley@techuk.org

Appendix

Glossary of terms

This glossary provides a quick reference for readers to understand key concepts discussed throughout the paper, ensuring clarity and consistency in the use of these terms for this paper's purposes.

AI, as defined by the OECD: An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment

AI Assurance, as defined by the Responsible Technology Adoption Unit[17]: 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.

Accountability, as defined by ISO/IEC 5723:2022[18]: The obligation of an organisation and its members to report, explain, and be answerable for the consequences of their AI-driven decisions and actions.

Audit, as defined by ISO[19]: Systematic, independent, documented process for obtaining records, statements of fact, or other relevant information and assessing them objectively, to determine the extent to which specified requirements are fulfilled.

Fairness metric, as defined by Ninareh Mehrabi[20]: A quantification of unwanted bias in training data or models.

Governance, as defined by IAPP[21]: A system of laws, policies, frameworks, practices and processes at international, national and organisational levels. AI governance helps various stakeholders implement, manage and oversee and regulate the development, deployment and use of AI technology. It also helps manage associated risks to ensure AI aligns with stakeholders objectives, is developed and used responsibly and ethically, and complies with applicable legal and regulatory requirements.

Model Card, as defined by OECD[22]: A brief document that discloses information about an AI model, like explanations about intended use, performance metrics and benchmarked evaluation in various conditions, such as across different cultures, demographics or race.

Safety, as defined by ISO/IEC 5723:2022[23]: Property of a system such that it does not, under defined conditions, lead to a state which human life, health, property, or the environment is endangered; safety involved reducing both the probability of expected harms and the possibility of unexpected harms.

Transparency, as defined by ISO/IEC 22989:2022[24]: Property of an organisation where appropriate activities and decisions are communicated to relevant stakeholders (3.5.13) in a comprehensive, accessible and understandable manner. Inappropriate communication of activities and decisions can violate security, privacy or confidentiality requirements.

References

1. GOV.UK. (2023). A pro-innovation approach to AI regulation. [online] Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#fn:178 [Accessed 17 Oct. 2024].
2. 14:00-17:00 (n.d.). ISO/IEC DIS 24029-2. [online] ISO. Available at: https://www.iso.org/standard/79804.html
3. ISO (2022). ISO/IEC 27001 standard – information security management systems. [online] ISO. Available at: https://www.iso.org/standard/27001
4. 14:00-17:00 (n.d.). ISO/IEC CD 5259-1. [online] ISO. Available at: https://www.iso.org/standard/81088.html
5. 14:00-17:00 (n.d.). ISO/IEC CD 5259-3. [online] ISO. Available at: https://www.iso.org/standard/81092.html
6. 14:00-17:00 (n.d.). ISO/IEC CD 5259-4. [online] ISO. Available at: https://www.iso.org/standard/81093.html
7. 14:00-17:00 (n.d.). ISO/IEC CD TR 5469. [online] ISO. Available at: https://www.iso.org/standard/81283.html
8. NIST (2021). AI Risk Management Framework. [online] NIST. Available at: https://www.nist.gov/itl/ai-risk-management-framework
9. IEEE Standards Association. (n.d.). IEEE Standards Association. [online] Available at: https://standards.ieee.org/ieee/7001/6929/
10. UK Government (2023). A pro-innovation Approach to AI Regulation. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper
11. 14:00-17:00 (n.d.). ISO/IEC TR 24027:2021. [online] ISO. Available at: https://www.iso.org/standard/77607.html
12. 14:00-17:00 (n.d.). ISO/IEC TR 24368:2022. [online] ISO. Available at: https://www.iso.org/standard/78507.html
13. 14:00-17:00 (n.d.). ISO/IEC 23894:2023. [online] ISO. Available at: https://www.iso.org/standard/77304.html
14. 14:00-17:00 (n.d.). ISO/IEC DIS 42001. [online] ISO. Available at: https://www.iso.org/standard/81230.html
15. 14:00-17:00 (n.d.). ISO/IEC CD TR 5469. [online] ISO. Available at: https://www.iso.org/standard/81283.html
16. 14:00-17:00 (n.d.). ISO/IEC 25059:2023. [online] ISO. Available at: https://www.iso.org/standard/80655.html
17. Introduction to AI assurance Ministerial foreword 1. Executive summary 2. AI assurance in context 3. The AI assurance toolkit 4. AI assurance in practice 5. Key actions for organisations 6. Additional resources. (2024). Available at: https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf
18. for, O. (2022). ISO/IEC TS 5723:2022. [online] ISO. Available at: https://www.iso.org/standard/81608.html [Accessed 17 Oct. 2024].
19. Conformity assessment techniques -Auditing. (n.d.). Available at: https://casco.iso.org/files/live/sites/cascoregulators/files/PDF/Annex%204%20-%20Conformity%20assessment%20techniques%20-%20Auditing.pdf [Accessed 17 Oct. 2024].
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), pp.1–35. doi: https://doi.org/10.1145/3457607
21. Anon, (n.d.). Key Terms for AI Governance. [online] Available at: https://iapp.org/resources/article/key-terms-for-ai-governance/
22. Oecd.ai. (2022). Model Cards - OECD.AI. [online] Available at: https://oecd.ai/en/catalogue/tools/model-cards [Accessed 17 Oct. 2024].
23. Iso.org. (2024). Available at: https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en:term:3.2.9 [Accessed 17 Oct. 2024].
24. Iso.org. (2024). Available at: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:12792:dis:ed-1:v1:en [Accessed 17 Oct. 2024].

techUK
FOR WHAT COMES NEXT

in linkedin.com/company/techuk

𝕏 @techUK

▶ youtube.com/user/techUKViews

✉ info@techuk.org

Image credits | iStock by Getty Images