

WHITE PAPER

# **Trustworthiness of Artificial** Intelligence

Recommendations of the TIC sector



TIC Council is the global trade association representing the independent TIC (Testing, Inspection and Certification) industry with over 100 members from leading Conformity Assessment Bodies (CABs) in 160 countries across the world. The TIC industry offers its services worldwide to ensure that only safe and secure consumer IoT products enter the market, to the benefit of consumers and the digital ecosystem worldwide.

+32 2 880 21 37

secretariat@tic-council.org

www.tic-council.org

## Abstract

From the gradual integration of artificial intelligence (AI) technologies into products and systems to the public availability of ChatGPT, artificial intelligence has gained considerable visibility among laypeople.

For AI developers and deployers, however, the AI landscape has become increasingly complex from a technical and regulatory perspective, creating gaps and overlaps that need to be addressed. Indeed, despite the existence of public and private mechanisms of hard and soft law, the AI ecosystem still lacks a common vision and a harmonised framework for designing and governing trustworthy AI models.

The TIC Council, the global trade association representing leading thirdparty Testing, Inspection and Certification (TIC) companies, is well-aware of this dual nature. While our members offer ever-efficient testing activities powered by AI technologies, they have also acquired expertise on testing AI products themselves.

The TIC industry can play a pivotal role between standards bodies, IT manufacturers, national and international decision-making bodies, and consumers. This Paper aims to provide a comprehensive overview of the TIC Council's vision for a trustworthy AI, based on the existing AI frameworks and a scientific discussion of the challenges ahead.

# Contents

Setting the stage for Trust in AI	1
The role and responsibilities of TIC in "Trust in AI"	3
Definition of Trustworthiness of AI	5
Challenges in the AI space	8
TIC Council Recommendations	10
Conclusion	12



### Setting the stage for Trust in AI

Artificial intelligence (AI) is currently reshaping the technological landscape. Its rapid advancement and integration into various sectors present both unparalleled opportunities and significant challenges. Breakthroughs ranging from autonomous vehicles to personalised healthcare systems are making headlines and are generating significant business value. However, this potential comes with an array of complexities that demand careful consideration, proactive solutions, and a global approach, particularly concerning trust effective mitigation of risks.

The AI ecosystem is characterised by its dynamic and fast-evolving nature, engaging a wide array of stakeholders from developers to end-users and regulators. The allure of AI technologies lies not only in the currently demonstrated capabilities but also in their promise to cater to the growing demands of a digitally connected world. Yet, this promise is accompanied by significant challenges that pose questions about the technology's reliability, security, and ethical implications. The lack of explainability of AI results, where even developers may struggle to decipher how AI models arrive at certain decisions, which can hinder trust and understanding of AI, underscores the need for transparent and accountable systems. Additionally, the diversity in ethical values among AI developers, influenced by their regional, cultural, and educational backgrounds, further complicates the landscape, suggesting a critical need for a common ground in ethical AI development. Moreover, a critical challenge arises from the insufficient cybersecurity expertise among AI developers, leaving AI systems vulnerable to potential threats and breaches. As AI becomes increasingly integrated into various domains, the need for robust cybersecurity controls is paramount to safeguard sensitive data and ensure the integrity of AI applications. These challenges are recognised for example in the standardisation roadmap by done by the German Institute for Standardisation (Deutsches Institut für Normung) aimed at identifying the gaps in standardisation and to inform policymakers.

The imminent rollout of regulations, such as the European Union's AI Act, highlights the growing recognition of AI's impact on society and the necessity for a structured framework to govern its development and application.

These regulations aim to mitigate systemic risks and ensure that AI systems are developed and used within ethical and secure boundaries. However, AI regulations vary across different countries and regions and a harmonised evaluation of AI systems poses significant challenges due to the lack of established standards and unified methods for assessment.

While there are several standards available or in the process of being published, the gap lies in operationalising these standards effectively. Without a track record demonstrating the efficacy of proposed requirements, the absence of benchmarks and agreed-upon conditions of acceptability complicates the evaluation process.

Addressing the rapid advancement of technology and the emergence of new machine learning models and generative AI poses unique challenges as these systems may lack specific use cases within scope for testing. Testing against undefined or unpredictable outputs generated by generative AI requires innovative approaches that go beyond traditional evaluation methods. In this context, there is a growing recognition that the TIC industry is a crucial component in the 'chain of trust' for AI. While acknowledging that risks may persist even with involvement from the TIC industry, the overall risk profile is significantly reduced compared to scenarios where such oversight is lacking. In this context, there is a growing recognition that the TIC industry is a crucial component in the 'chain of trust' for AI. While acknowledging that risks may persist even with involvement from the TIC industry, the overall risk profile is significantly reduced compared to scenarios where such oversight is lacking. In this context, there is a growing recognition that the TIC industry is a crucial component in the 'chain of trust' for AI. While acknowledging that risks may persist even with involvement from the TIC industry, the overall risk profile is significantly reduced compared to scenarios where such oversight is lacking.

The TIC industry boasts over a century of experience in evaluating (new) technologies embodying core values of integrity and impartiality. The TIC industry is uniquely positioned to address the multifaceted challenges of AI. Its role in the 'chain of trust' for AI is pivotal, not only in ensuring compliance with upcoming regulations but also in embodying the principles of trustworthy AI, including transparency, robustness, reliability, and controllability.

TIC industry, with its rich heritage and commitment to upholding the highest standards, stands ready to lead these efforts. By leveraging its expertise and values, the TIC industry not only supports the assurance and certification of Al systems but also advocates for the creation of a trustworthy Al ecosystem that is conducive to nurturing innovation and aligns with societal values and expectations.

# The role and responsibilities of TIC in "Trust in AI"

Different approaches to regulating AI are currently being pursued in many countries and regions around the world. These approaches range from voluntary guidelines and codes of conduct to binding legal regulations. In countries such as the USA, the UK, Japan, Singapore, Korea and China, there are a number of regulatory measures that are not yet legally binding. What most approaches have in common is that they generally take a risk-based approach to AI. Rather than regulating a technology, these approaches can be understood as regulating an outcome. This risk-based approach attempts to strike a healthy balance between mitigating risk and promoting AI innovations. Leading examples include the EU AI Act or the Canadian AI and Data Act, both of which propose risk and impact assessments to categorise AI systems into compliance obligations.

### The Many Types of AI Governance



Source: Putting Standards Into Action - Responsible AI Institute

The AI landscape will continue to develop rapidly on a global level and will therefore require a harmonised approach in the future with standardised compliance requirements and testing standards, particularly for high-risk AI systems. The TIC industry plays a crucial role in this context, as it has the necessary experience from the conformity assessment procedures of existing technologies as well as tools, competences, and qualifications to identify and independently assess AI risks.

For establishing trust in AI, the TIC industry contributes the following:

### Familiarity with Regulatory Requirements

TIC Council members are well-versed with the regulatory requirements of different regions. Their services provide assurance of compliance for a functioning market. This includes meeting of regional and sectorial legislative provisions that intersect with AI, such as the application of smart robotics in the manufacturing industry, the testing and certification of safety-critical products including medical devices, and AI enabled autonomous driving. It has to be noted that all AI regulations have to be considered in addition to existing legal requirements such as cybersecurity, data privacy, and product safety provisions.

#### **AI Assessment Considerations**

When assessing AI, it is important to consider whether a product or an industrial asset has already undergone a sectoral conformity assessment involving a notified third-party body. This could impact the scope and depth of testing. Factors to consider include product safety regulation, cyber, ethics, ESG and AI aspects of products, and a holistic third-party conformity assessment approach when mandated by different legislations. It is important though, that for each applicant of the product or asset, that will be put on in the market, a holistic third-party conformity assessment approach is striven for and at least the scope of the assessments and the underlying test areas were clearly documented.

#### **Maintaining and Developing Competences**

The members of the TIC Council are obliged to continuously maintain and further develop their competences and qualifications in view of the state of technical development. A regular exchange of experience focusing on the parameters for the implementation of appropriate conformity assessment procedures according to the current state of the art, suitable procedures and market observation procedures has therefore long been mandatory in the TIC industry.

#### Monitoring AI in the Market

Monitoring AI in the market involves developing experience, risk assessment experience, monitoring of certification marks, and monitoring of given and product application domains with regards to the defined intended as well as unintended use.

### **Active Involvement in Standardisation Work**

In addition, TIC companies have always been actively and transparently involved in standardisation work. Standards are created in a consensus-based process and provide guidance for meeting regulatory requirements. They have to address the technical challenges but remain practical applicable. The establishment of a suitable infrastructure for collecting and analysing relevant information would therefore be highly welcomed.

#### **Communication and Awareness**

Further the information and experience gained from these activities should also be used for future communications. This will help to raise public and consumer awareness of the important role of the TIC industry and the need to actively involve the TIC industry in the assessment of AI applications. It is important to specify that the involvement of third parties helps to identify and mitigate risks, but does not guarantee risk prevention, which is the responsibility of the producer or distributor of the product or application.

### **Definition of Trustworthiness of AI**

This section presents a common understanding of the aspects that constitute a trusted AI system. We delve areas that can be assessed by independent third parties to strengthen the credibility of AI systems. Further information about AI terminology can be found here.

### **Understanding Trustworthiness in AI**

Trust in AI is not a singular, easily defined concept but rather a multifaceted one. It encompasses a range of elements that must be appropriately addressed to ensure a system's reliability and societal acceptance. While there is no universal definition of trustworthy AI, several prominent organisations, such as the European Union (EU), National Institute of Standards and Technology (NIST), and Organisation for Economic Co-operation and Development (OECD), have made significant strides in framing what constitutes trust in AI. Despite the fragmented efforts across these organisations, there exists a consensus on the key aspects of an AI system, which have been crucial in developing standards like ISO 22989, titled "Artificial Intelligence – Concepts and Terminology," which introduces key concepts and terminology.

#### **Consensus and Key Aspects**

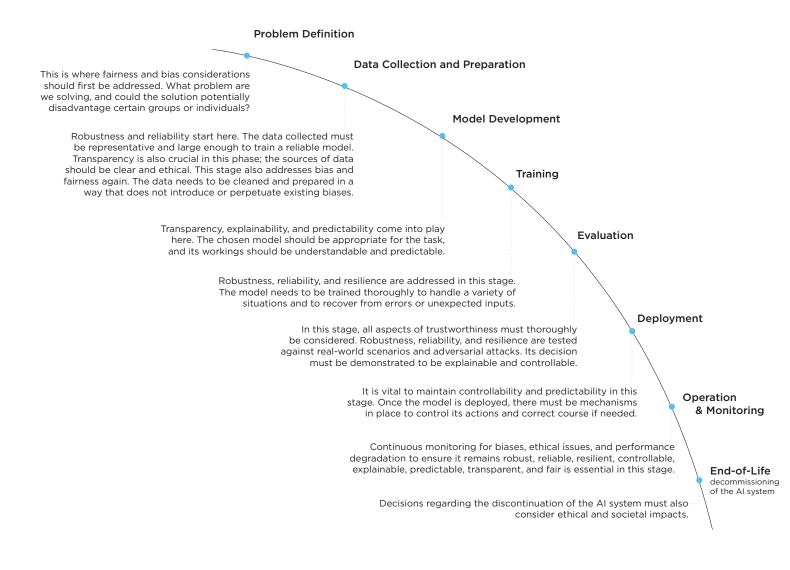
Several key aspects of AI have been defined in ISO 22989: Robustness, Reliability, Resilience, Controllability, Explainability, Predictability, Transparency, and Bias and Fairness. These aspects collectively contribute to the trustworthiness of AI systems. Additionally, to the key aspects defined in ISO 22989, the integration of robust privacy measures is essential for trustworthiness of AI systems.

- **Robustness:** An AI system's ability to maintain performance under varying conditions and to withstand manipulative attacks or errors.
- **Reliability:** This pertains to the system's ability to perform its intended function consistently over time.
- **Resilience:** This aspect refers to the system's capacity to recover from adverse events and adapt to changes in the environment.
- **Controllability:** Ensuring that human operators can intervene and control the AI system at any given time is crucial for safety and ethical reasons.
- **Explainability:** The system's decisions and processes should be understandable to its users and developers. This transparency fosters trust and facilitates troubleshooting.
- **Predictability:** A degree of certainty in how the AI system behaves under different circumstances is essential for trust.
- **Transparency:** Openness about how the AI system works, its limitations, and the data it uses is vital for accountability.
- **Bias and Fairness:** This involves ensuring that the AI system does not perpetuate societal biases and is fair in its operations and outcomes.
- Privacy: An AI system's ability to ensure robust data protection and confidentiality, to maintain user trust and comply with ethical and legal standards.

ISO 22989 plays a pivotal role in standardising the understanding of these key aspects. It provides a framework for assessing AI systems, linking the commonly agreed-upon definitions and the terminologies used across different standards and guidelines. The ISO standard serves as a bridge, connecting the dots between various definitions and providing a common language for discussing AI trustworthiness.

### AI Life Cycle and Trustworthiness

Another crucial aspect of trustworthy AI is understanding its life cycle, which encompasses the stages from design and development to deployment and operation. Trustworthiness must be a consideration at each stage of this life cycle:



For a comprehensive evaluation of AI systems' trustworthiness, a management system approach is necessary. This includes:

- **Governance:** Establishing policies and practices that ensure the AI system aligns with ethical and legal standards.
- **Risk Management:** Identifying and mitigating risks associated with AI applications, particularly in critical areas like healthcare or finance.
- **Quality Assurance:** Regularly assessing the system's performance and making necessary adjustments to maintain its trustworthiness.
- **Stakeholder Engagement:** Involving users, developers, ethicists, and the public in discussions about the AI system to ensure it meets societal needs and values.

Recently, a new standard for AI Management Systems, ISO/IEC 42001, was published, addressing these aspects.

In conclusion, the trustworthiness of AI systems is a multifaceted concept that requires a holistic approach. The common understanding, as derived from the guidelines of EU, NIST, OECD, and encapsulated in the ISO 22989, is the basis for assessing AI systems. By considering the aspects of robustness, reliability, resilience, controllability, explainability, predictability, transparency, privacy and bias and fairness throughout the AI life cycle, and incorporating a robust management system, we can ensure the development of AI systems that are not only technologically advanced but also ethically sound and socially acceptable.

### **Challenges in the AI space**

The TIC industry finds itself well positioned to be at the forefront of assessing the trustworthiness of AI systems, leveraging its expertise and industry engagement to ensure the safety, security, and ethical deployment of AI technologies. This section further explores the principles of trustworthy AI, the role of third-party testing, and related challenges. Finally, it makes some recommendations for the implementation trustworthy AI.

### **Basic Principles for Functional Trustworthiness**

Al assessment must not depend solely on document-and evidence-based reviews, which may not suffice for the dynamic and complex nature of Al systems. The TIC industry advocates for a more robust approach that includes:

- A precise definition of intended use.
- Collection of risk-based minimum performance requirements.
- Statistically valid system testing based on independent random samples.

These principles aim to elevate the functional trustworthiness of AI systems, ensuring they meet high standards of safety and efficacy. Additionally, the TIC industry advocates for robust security assessments of AI technologies that include:

- Red-Team testing simulating adversarial attacks, helping to identify vulnerabilities in AI systems that might not be apparent in conventional testing environments.
- Application of Safe and secure centric processes throughout the AI system's lifecycle.
- Cybersecurity evaluations focused on vulnerabilities in AI systems and their underlying software, including penetration testing, security audits, and the development of AI-driven tools to enhance cybersecurity measures.

There are several ongoing standardisation efforts that facilitate the creation of harmonised approaches to requirements and assessment for AI systems. These efforts are crucial for harmonising practices across borders and sectors, allowing for a consistent evaluation framework. However, the path towards comprehensive standardisation is ongoing, while many areas are still being explored.

There has been progress in key areas such as the development of ISO/IEC 42001, a standard that outlines requirements for establishing, implementing, maintaining, and continually improving an artificial intelligence management system within organisations. It aims to ensure the responsible development and use of AI systems by addressing ethical considerations and transparency. However, some practical challenges remain, including the need for sector-specific adaptations due to the broad applicability of ISO/IEC 42001 across various industries. Additionally, there is a necessity for ensuring continuous

alignment with evolving legal and regulatory landscapes. Furthermore, addressing the inherent complexities of fast-paced evolving AI technologies, especially in areas like autonomous decision-making and ongoing learning, poses significant challenges.

### Third-Party Assessment/Testing

Third-party assessment/testing is recognised for its potential to provide an unbiased evaluation and for AI systems, though this is universally true, it is particularly relevant for systems classified as high-risk. The feasibility of such assessments hinges on the availability of clear, standardised methodologies, tools, and assessment criteria and the expertise of the TIC industry in applying these criteria across various AI fields of application. This practice is in the process of progressing from "under academic research" towards "wellestablished practice," with varying degrees of adoption across different sectors.

### Challenges

The assessment, testing, and evaluation of AI systems pose unique challenges, including:

- **Complexity of AI Systems:** The intricate nature of AI algorithms, especially those non-deterministic and involving deep learning, makes it difficult to understand and predict their behaviour. This complexity challenges traditional testing methodologies.
- **Dynamic Nature of AI:** AI systems can learn and evolve over time, which means a system deemed safe at one point might develop unforeseen behaviours or vulnerabilities as it interacts with new data. Defined processes for ongoing testing and monitoring of AI systems are essential.
- Lack of Standards and Best Practices: While efforts towards standardisation are in progress, there is still a lack of universally accepted standards and best practices for the evaluation of many aspects of AI systems. This complicates the establishment of a consistent assessment framework and emphasises the need for knowledgeable and trusted third-party expert evaluators.

Despite these challenges, the TIC industry is actively involved in shaping the future of AI assessment through various activities. These include advancing methodologies, collaborating for standardisation, and promoting transparency and trust through independent third-party testing.

As AI continues to be deployed in various sectors, the role of the TIC industry in ensuring the trustworthiness of these systems becomes increasingly vital. By addressing the challenges and leveraging the opportunities outlined, the TIC industry can provide invaluable insights and guidance, setting the stage for recommendations and contributing to the safe, ethical, and effective deployment of AI technologies globally.

### **TIC Council Recommendations**

As nations around the world, including the European Union with its AI Act and the United States through its Executive Order on AI, embark on establishing comprehensive frameworks to govern the development and use of AI, the role of the TIC industry becomes increasingly critical in ensuring that AI systems are not only compliant with these diverse regulations but also meet the standards of safety, security, and ethical considerations.

Recognising the complexities and challenges posed by the varied international regulatory landscapes, this section outlines strategic recommendations from the TIC Council. These suggestions are designed to advocate for a unified approach that fosters trust, reliability, and innovation in AI technologies. The TIC Council's recommendations aim to bridge regulatory differences and reinforce the indispensable role of TIC companies in shaping a responsible and secure future for AI, without compromising the innovation and progress of the technology:

- 1. Adoption of Risk-Based AI Assessment Frameworks: Frameworks used to assess AI systems should classify them based on their potential impact and risks, such as risks to national security, health, safety, or fundamental rights, and economic risks, facilitating appropriate levels of scrutiny and based on factors such as the intended use of the system, the type of data used to train the system, and the intended operational environment. For this, there is a need for the development and implementation of thorough risk assessment frameworks specific to AI systems, that consider both inherent risks of AI technologies and external cybersecurity threats.
- 2. Robust Security-by-Design Principles: Safety and security-by-design principles should be incorporated in AI system development. Robust security measures should be integrated from the initial design phase and different types of testing included through the entire lifecycle of the system. This approach includes proactive security integration, data quality assessment, continuous safety and security assessments, stringent data protection, resilience against cyberattacks, transparency in operations, and collaborative efforts among stakeholders. Such a strategy ensures that AI systems are functional, secure, resilient, and trustworthy, effectively addressing potential risks and vulnerabilities when they are identified.
- **3. Development Comprehensive AI Testing Framework:** Given the diverse applications and impacts of AI, The TIC Council recommends the development of an 'AI Testing Framework' or 'AI Certification Framework' (AICF). This framework should be modular, allowing for different 'Building Blocks' corresponding to various dimensions of trustworthy AI, such as robustness, reliability, cybersecurity, bias, fairness, and transparency. This modular approach would enable tailored assessments depending on the specific context and application of the AI system. An AI Testing Framework of AICF would need to be developed, incorporating clear testing methods and evaluation criteria that ensure interoperability between testing labs. There already existing important framework on

the topic, on which future work could rely, such as the NIST AI Risk Management Framework.

- 4. Functional guarantees and verification of AI systems: The TIC Council acknowledges the pioneering efforts of the EU AI Act but identifies areas for enhancement. Specifically, there is a need for more clearly defined, functionally verifiable requirements for AI systems. Conformity assessment procedures should not be based extensively on documentation without sufficient emphasis on functional guarantees, especially in safety-critical applications, may not be adequate to ensure the trustworthiness of AI systems. We urge for a balance between documentation requirements and practical, functional assessments which include both data and AI models.
- **5.** Independent Conformity Assessment: AI Systems should undergo independent conformity assessments by qualified 3rd parties, particularly High-risk AI systems. This ensures that these systems adhere to the standards of safety and security; the independent nature of third-party assessments helps to ensure that unforeseen risks are correctly identified, especially for new and emerging technologies.
- 6. Lifecycle Management of AI Systems: Incorporating security, continuous monitoring of the reliability and fairness of the system, as well as reassessment throughout their lifecycle is recommended. Many AI systems are not static; changes can occur due to software updates, learning from new data, or alterations in the operating environment. Without regular assessments, an initially safe and compliant AI system could become unsafe, unsecure and/or non-compliant. Lifecycle management processes that include continuous monitoring helps to make certain that such issues are identified and rectified promptly, maintaining the integrity and trustworthiness of the system. Implementation of an AI Management Systems is strongly recommended to ensure implementation of the correct processes and policies to develop, manage and operate AI systems throughout its entire lifecycle.
- 7. Horizontal vs sectorial regulation and standardisation: TIC Council acknowledges the efforts from the EU to come up with regulation and standards as horizontal as possible, covering all industries under the same legal umbrella. Still, differences among different industries should be considered, acknowledging, and respecting the sectorial particularities which need to be translated into adaptations of the standards.

The TIC industry possesses a wealth of experience, and not just in testing and certification, positioning it as an invaluable and trusted partner throughout every stage of a product's lifecycle. This expertise is particularly crucial for organizations aiming to navigate the intricate and ever-changing landscape of global regulatory requirements. With its comprehensive understanding and adaptability, the TIC industry is a key ally with multi-domain experience in achieving success in complex regulatory environments.

### Conclusion

Since over 150 years and starting in the first industrial revolution, the TIC industry finds itself well positioned to be at the forefront of assessing new technologies. With the trustworthiness of AI systems, TIC industry is leveraging its expertise and industry engagement to ensure the safety, security, and ethical deployment of AI technologies.

As we have previously seen a global approach and extensive experience in conformity assessment is useful and necessary for safeguarding new technologies such as Al-based products and services. This is particularly true as these products and services are subject to increasingly complex compliance requirements that generally relate to more than just one or a few legal acts, applicable technical norms and standards or other regulatory works.

In this respect, it is important for active risk management and the safeguarding of AI-based applications that Conformity Assessment Bodies are entrusted with the conformity assessment, which can both ensure familiarity with regulatory requirements and ensure proven procedures and processes for maintaining and developing competences. In addition, active involvement in standardisation works as well as appropriate monitoring of AI in the market contribute to the dynamic development of risks in innovative technologies regarding necessary testing requirements throughout the entire AI life cycle and thus to being able to adequately assess the trustworthiness of these applications.

Although the design and definition of the basic principles for the assessment of AI applications is currently still in the middle of being developed in the standardisation committees, we already have sufficient knowledge of important cornerstones for the testing and certification of these applications from the perspective of TIC industry members with experience in conformity assessments.

Therefore, we suggest a unified approach that fosters trust, reliability, and innovation in AI technologies. The TIC Council's recommendations aim to bridge regulatory differences and reinforce the indispensable role of TIC companies in shaping a responsible and secure future for AI, without compromising the innovation and progress of the technology.

Al assessment must not depend solely on document-and evidence-based reviews, which may not suffice for the dynamic and complex nature of Al systems. The TIC industry advocates for a more robust approach that includes:

- A precise definition of intended use.
- Collection of risk-based minimum performance requirements.
- Statistically valid system testing based on independent random samples.

These principles aim to elevate the functional trustworthiness of AI systems, ensuring they meet high standards of safety and efficacy. Additionally, the TIC industry advocates for robust security assessments of AI technologies. Thus, we see as the basic principles for Functional Trustworthiness.

Furthermore, we believe that the following guardrails will support to bridge the gap between innovative power and progress of technology and regulatory constraints towards a pro-active risk management:

- Adoption of Risk-Based AI Assessment Frameworks
- Robust Security-by-Design Principles
- Development Comprehensive AI Testing Framework
- Functional guarantees and verification of AI systems
- Independent Conformity Assessment
- Lifecycle Management of AI Systems
- Respecting differences in horizontal vs sectorial regulation and standardisation

With regards to these principles, guardrails and proven assessment procedures and competences of the TIC industry will support the efforts to make the future development and operations of AI technology as safe as possible and ensure that AI offers great opportunities to society.



### Editor's Note About TIC Council

TIC Council is the global trade association representing the independent third-party Testing, Inspection and Certification (TIC) industry which brings together about 100-member companies and organizations from around the world to speak with one voice. Its members provide services across a wide range of sectors: consumer products, medical devices, petroleum, mining and metals, food, and agriculture among others. Through provision of these services, TIC Council members assure that not only regulatory requirements are met, but also that reliability, economic value, and sustainability are enhanced. TIC Council's members are present in more than 160 countries and the wider TIC sector currently employs more than 1 million people across the globe.

TIC Council Secretariat Rue du Commerce 20-22 B-1000 Brussels, Belgium secretariat@tic-council.org www.tic-council.org