Non-regulatory Approaches to AI Governance: Spotlight on AI Assurance Techniques and Standards

EAA e-Conference on
Data Science & Data Ethics

16 May 2023

Nuala Polo

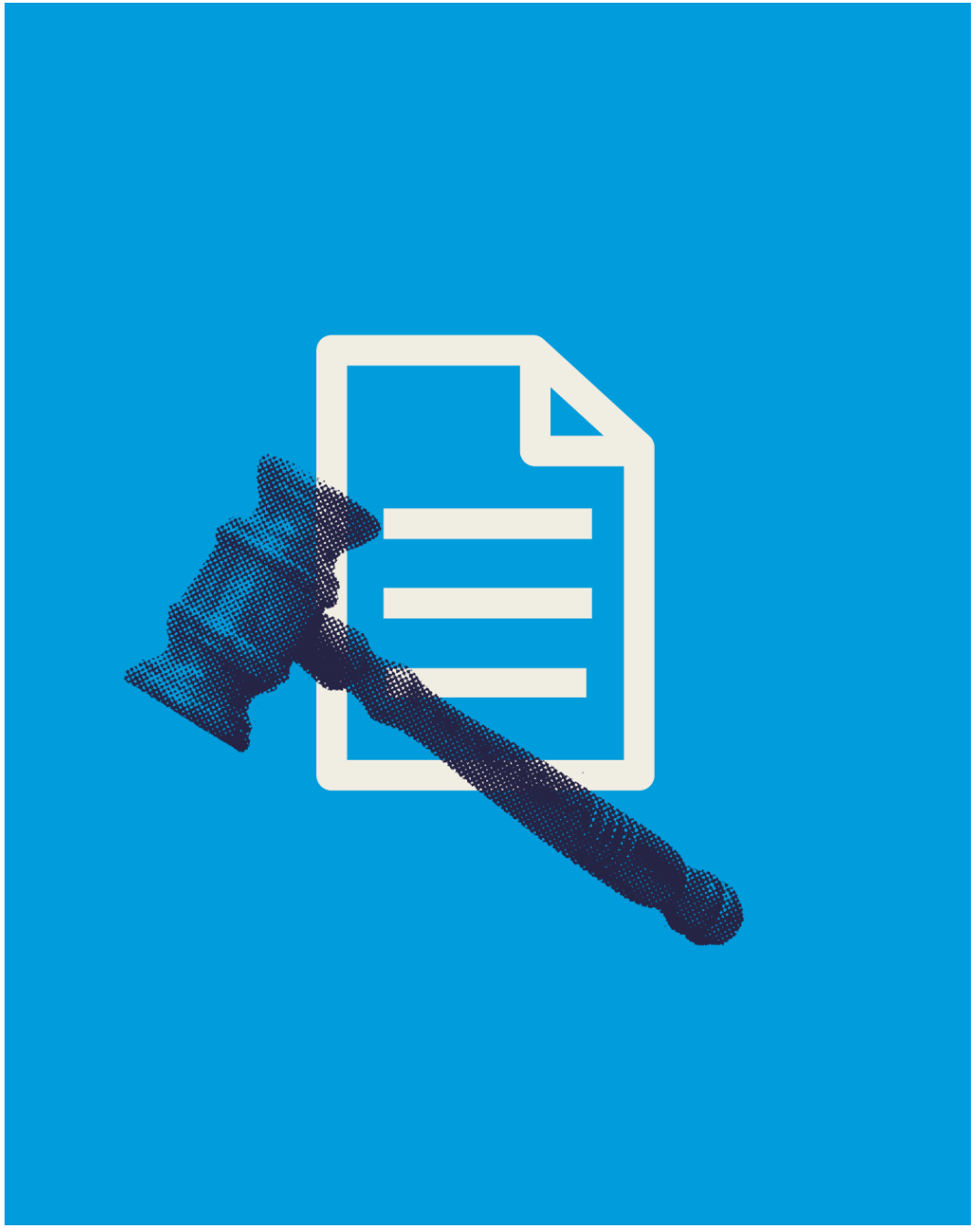
Senior Policy Advisor at the UK's Centre for Data Ethics and Innovation (CDEI)

european actuarial academy

AGENDA

1. Overview of AI governance landscape
2. Introduction to non-regulatory tools for trustworthy AI
3. Deep dive: AI assurance techniques
4. Deep dive: Standards
5. UK State of play
6. Resources and next steps

1. OVERVIEW OF AI GOVERNANCE LANDSCAPE



OVERVIEW OF AI GOVERNANCE

AI Governance refers to mechanisms including **laws, regulations, policies, institutions, and norms** that set out processes for making decisions about AI.

The goal of AI governance is to maximising the benefits, while mitigating potential risks and harms.

Key elements of our pro-innovation framework

Cross-sectoral principles

Our framework will be underpinned by a set of cross-sectoral principles including concepts such as transparency, safety, and security, to guide how actors in the AI ecosystem approach responsible AI and AI risk.

Based on OECD Principles

Leveraging existing regulator expertise

We will leverage the sector expertise of our world-class regulators, focusing on outcomes rather than the technology itself. We balance the economic and societal potential benefits of AI against its risks.

Context-specific

We acknowledge that AI is a dynamic, general purpose technology and that the risks arising from it depend principally on the context of its application. The same AI used in different contexts may need regulating differently.

Central functions to drive coherence

To ensure that the overall framework offers a proportionate, coherent and effective response to risk while promoting innovation across the regulatory landscape.

5

The proposed cross-cutting principles

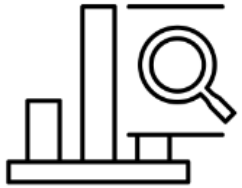
Existing regulators will be expected to implement the framework underpinned by 5 values-focused cross-sectoral principles, based on the **OECD AI Principles**

Safety, Security & Robustness	AI systems should function in a robust, secure and safe way throughout the AI life cycle, and risks should be continually identified, assessed and managed.
Appropriate Transparency & Explainability	AI systems should be appropriately transparent and explainable.
Fairness	AI systems should not undermine the legal rights of individuals or organisations, discriminate unfairly against individuals or create unfair market outcomes. Actors involved in all stages of the AI life cycle should consider definitions of fairness that are appropriate to a system's use, outcomes and the application of relevant law.
Accountability & Governance	Governance measures should be in place to ensure effective oversight of the supply and use of AI systems, with clear lines of accountability established across the AI life cycle. AI life cycle actors should take steps to consider, incorporate and adhere to the principles and introduce measures necessary for the effective implementation of the principles at all stages of the AI life cycle.
Contestability & Redress	Where appropriate, users, impacted third parties and actors in the AI life cycle should be able to contest an AI decision or outcome that is harmful or creates material risk of harm.

Central functions

Monitoring and evaluating the regulatory framework

Identifying, assessing and monitoring AI risks



Conducting horizon scanning and gap analysis

Supporting testbeds and sandbox initiatives

Providing education and awareness

Promoting interoperability with international frameworks

These functions require and support **collaboration** between government and a range of stakeholders:

- Regulators
- International partners
- Industry
- Civil society
- Academia
- Public

INTERNATIONAL APPROACHES

International approaches to AI Governance

UK
AI Regulation: A Pro-Innovation Approach

EU
EU AI Act
AI Liability Directive

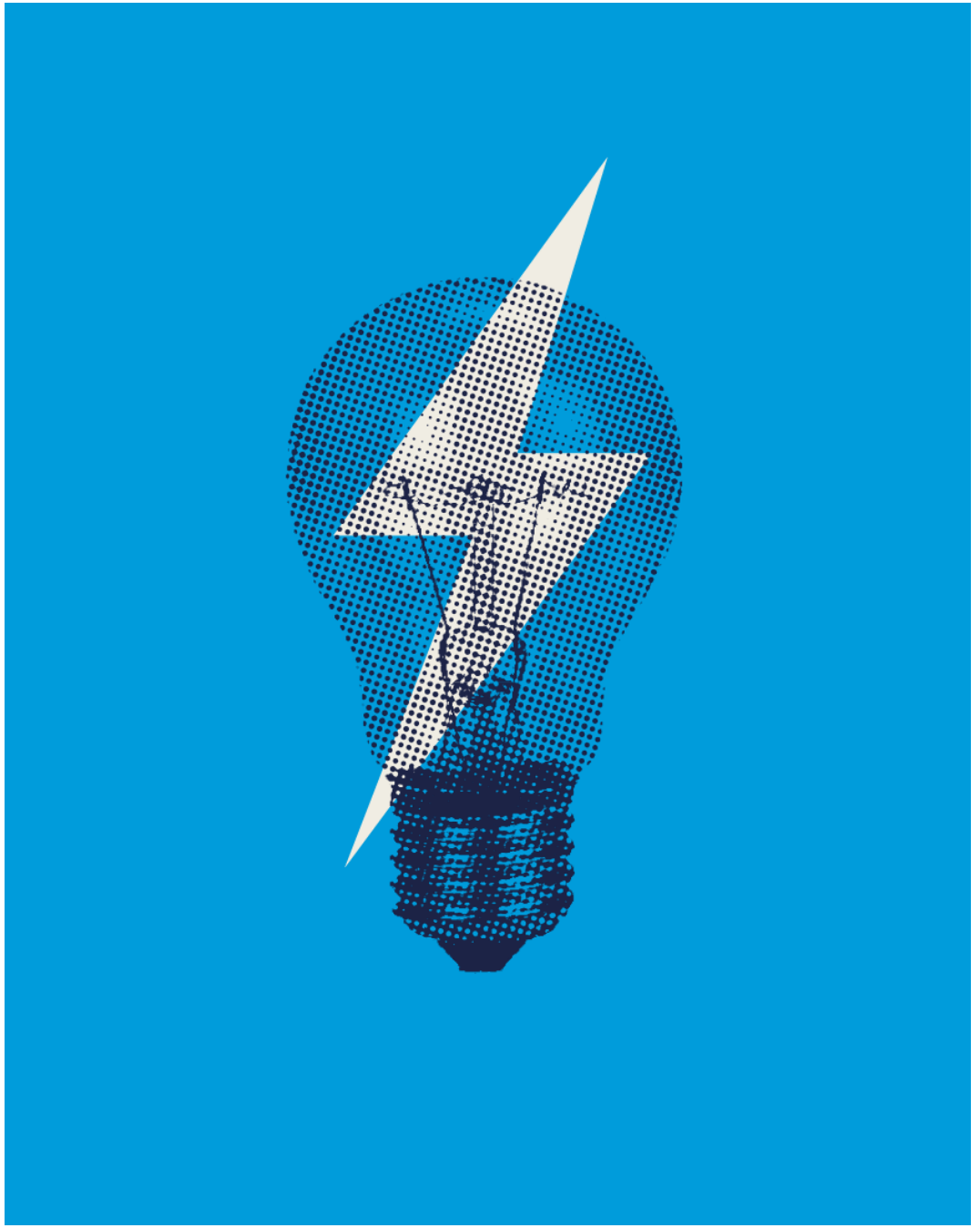
Canada
Directive on Automated Decision Making
AIDA

US
Blueprint for an AI Bill of Rights (BOR)
NIST RMF

China
MIST AI ethics principles
CAC secondary legislation

Source: CEIMIA "A Comparative Framework for AI Regulatory Policy", February 2023.

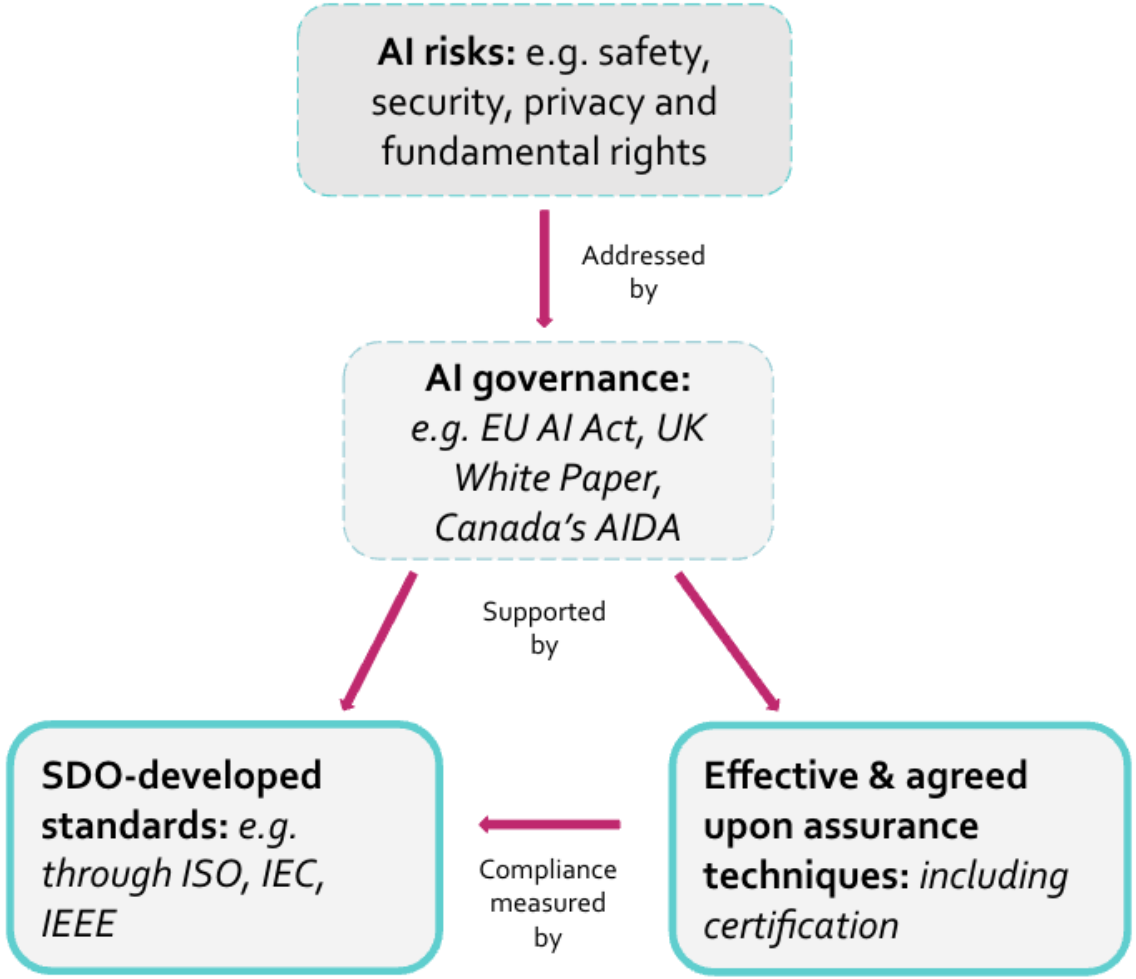
2. INTRODUCTION TO NON-REGULATORY TOOLS FOR TRUSTWORTHY AI



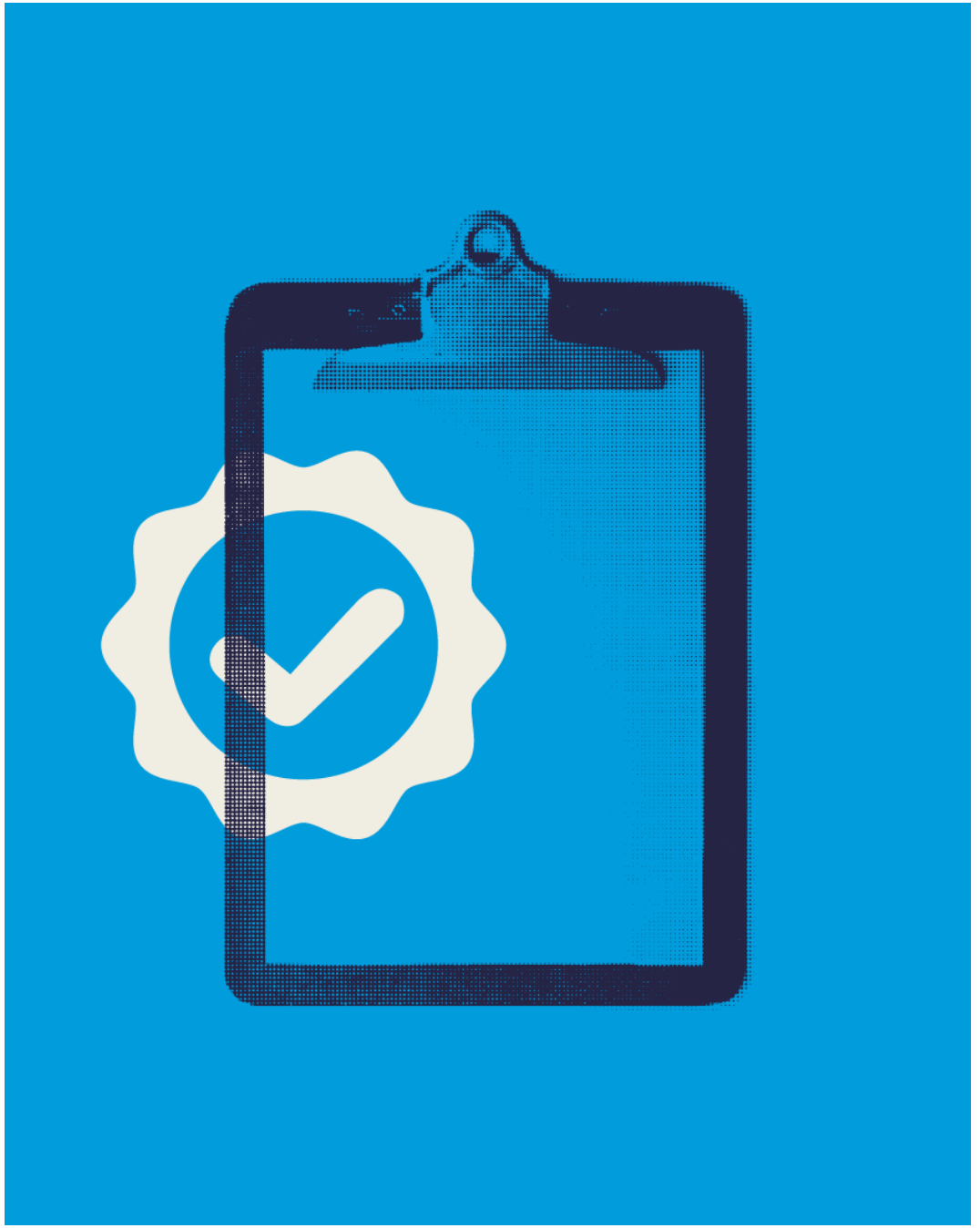
TOOLS FOR TRUSTWORTHY AI

Tools for trustworthy AI will play a critical role in enabling the responsible adoption of AI by **supporting the implementation of regulatory framework** and **boosting international interoperability.** These include:

1. Assurance techniques
2. Consensus-based standards

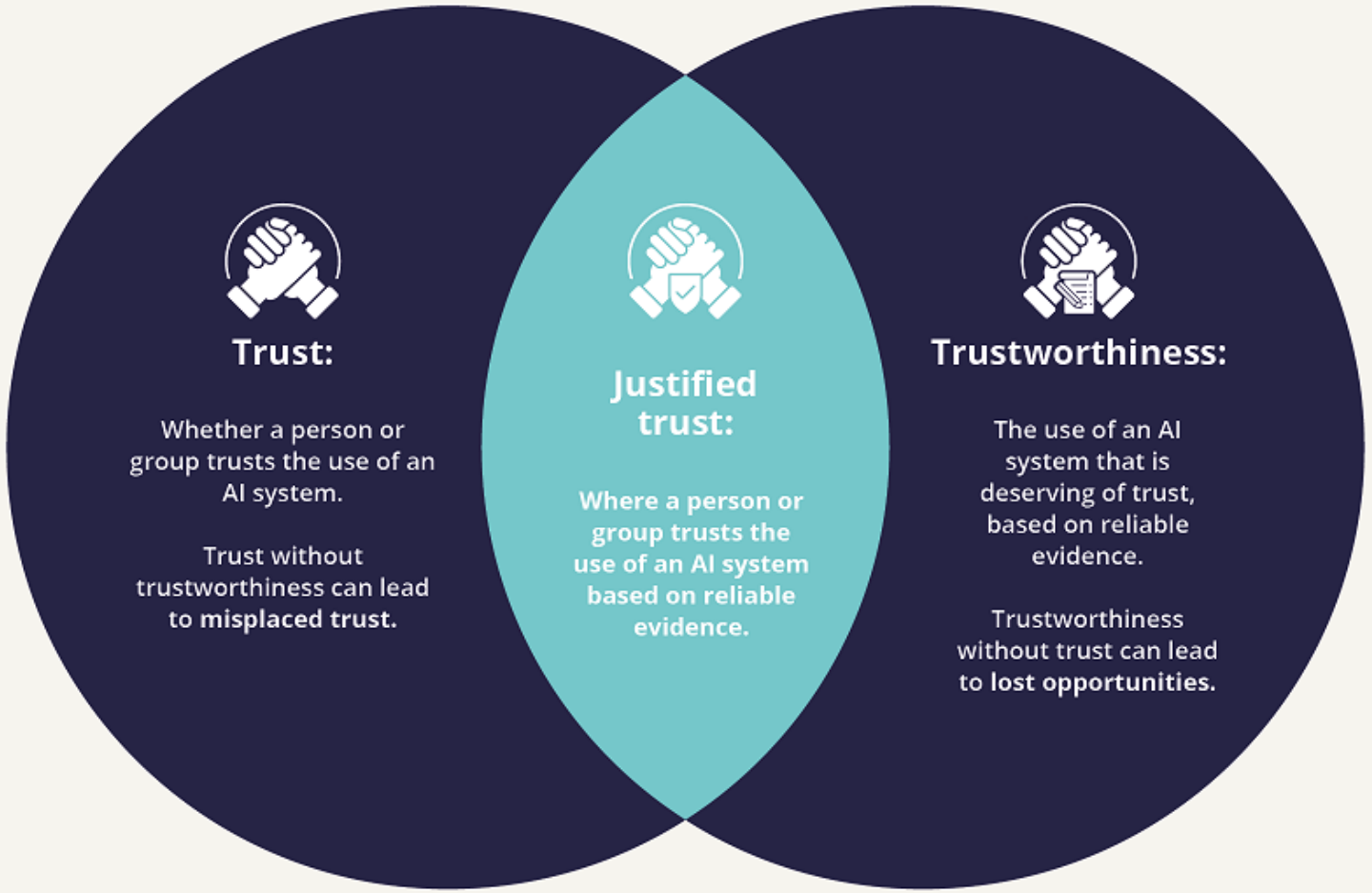


3. DEEP DIVE: ASSURANCE TECHNIQUES



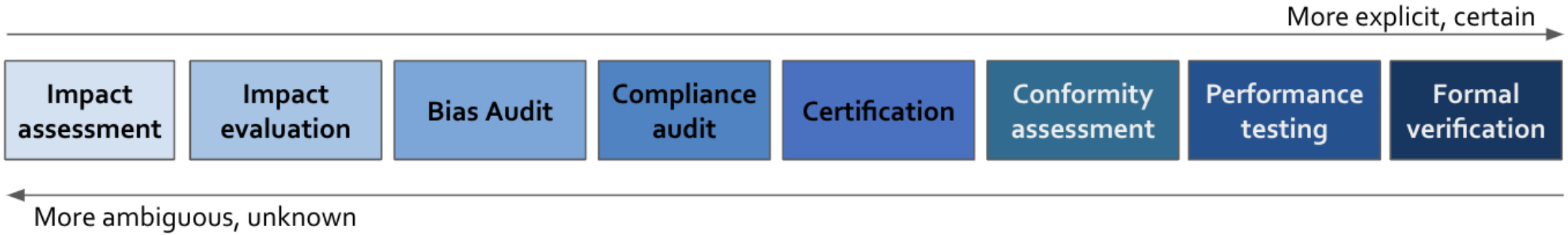
AI ASSURANCE TECHNIQUES

The relationship between trust, trustworthiness and justified trust

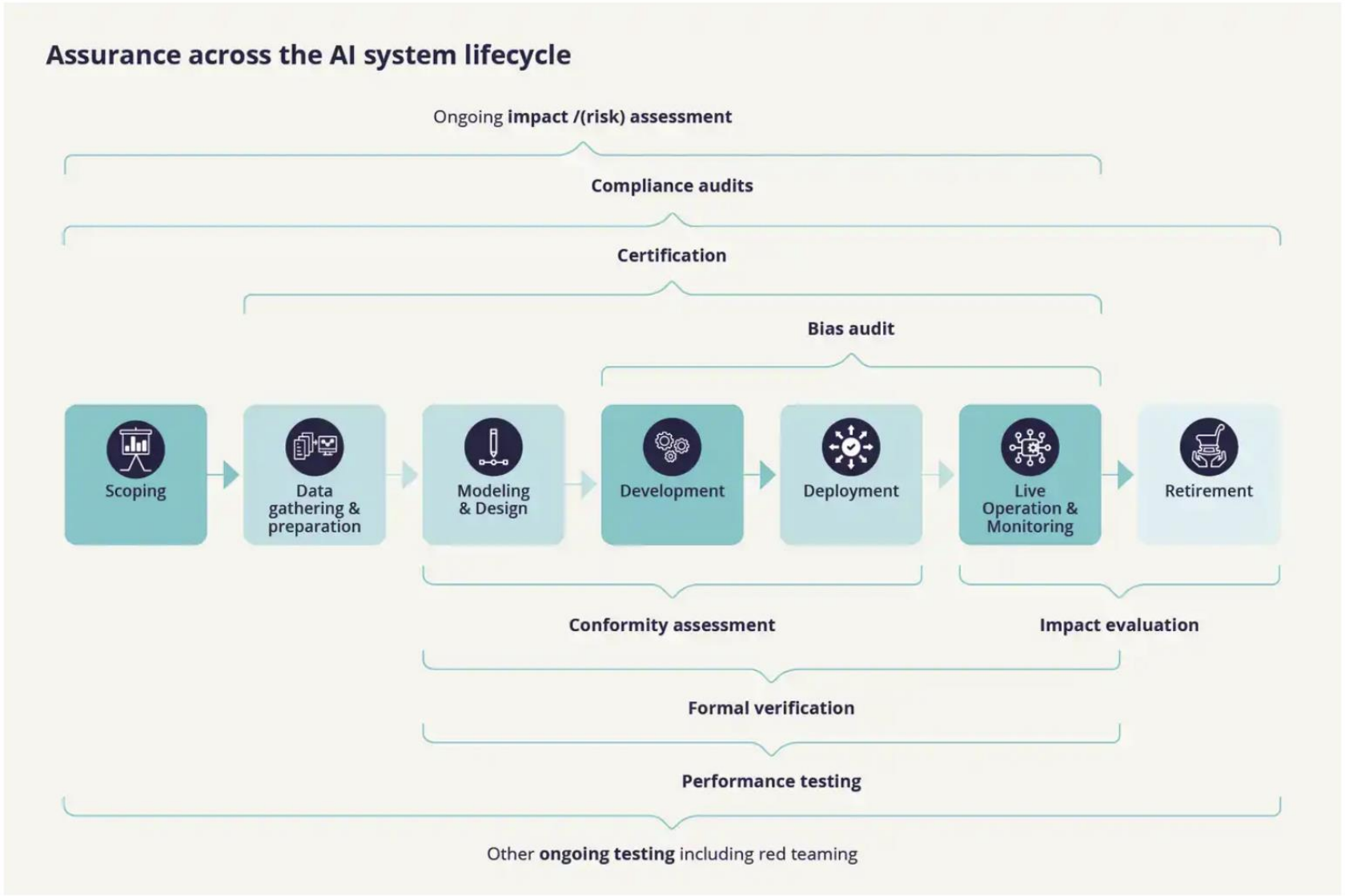


AI ASSURANCE TECHNIQUES

- The goal of AI assurance techniques, are to **measure, evaluate and communicate whether** AI systems are trustworthy.
- There are a **range of different techniques** for assuring AI systems.
- The challenge is to adopt a combination of, or **toolkit of assurance methods**, where the right tool is adopted for the right kinds of subject matter.



AI ASSURANCE TECHNIQUES



4. DEEP DIVE: STANDARDS

TECHNICAL STANDARDS

Standards provide a reliable basis for people to **share the same expectations** about a product, process system or service.

Standards...

- ...are mostly developed by **industry** in Standards Development Organisations (SDOs), such as ISO or ETSI
- **...are voluntary**
- ...are usually formal documents that establish uniform engineering or technical criteria, methods, processes, and practices

TECHNICAL STANDARDS

There are many different types of standards, including:

Foundational standards build common understanding around definitions and terminology

Process standards universalise best practice in organisational management and governance

Measurement standards define metrics and methods for quantitative measurement

Performance standards set specific performance thresholds for acceptability

Without standards of some kind, we have **advice**, rather than **assurance**.

5. UK STATE OF PLAY

UK STATE OF PLAY

- The CDEI's "Roadmap to an Effective Assurance Ecosystem" outlined the role of assurance mechanisms in mitigating the potential risks of AI, and maximising the benefits of these systems.

- In theory, AI assurance can help to build justified trust in AI systems. But **what does this look like in practice**?

- Following publication, we engaged with industry stakeholders to understand current levels of awareness of, and engagement with AI assurance.

UK STATE OF PLAY

The *Industry temperature check: Barriers and enablers to AI assurance* was developed by analysing data from **four industry engagement activities.**

We conducted **qualitative thematic analysis** on the data collected from each of these activities.

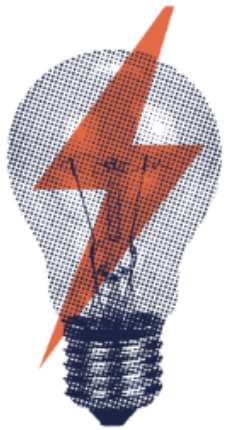
This report summarises key findings from across our engagements.

- **Ministerial Roundtables**
- **Semi-structured interviews**
- **CDEI x techUK AI assurance symposium**
- **Targeted online survey**

GENERAL FINDINGS: BARRIERS

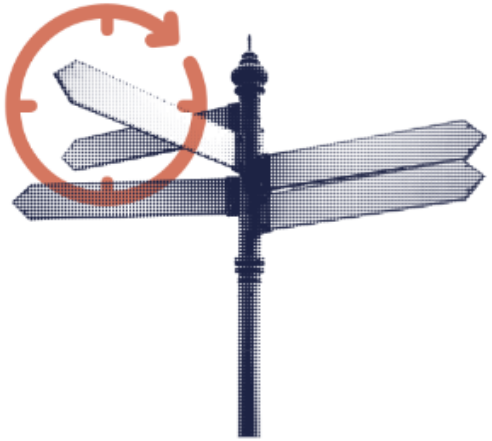
General findings: Barriers to engaging with AI Assurance across sectors

Lack of knowledge/skills



Lack of buy-in/demand

Lack of resources



Lack of standardised approach

Regulatory uncertainty

GENERAL FINDINGS: INTERVENTIONS

1.

Repositories and guidance

- Desire for **"concrete and operational guidance"** to aid in the identification of assurance techniques and standards
- Standards → Recently launched AI Standards Hub
- Appetite for similar repositories of AI assurance **techniques**

2.

Support for SMEs

- SMEs reported having **limited resources and expertise** to implement AI assurance
- Desire for a **library of free tools** to support SMEs identify/use AI assurance techniques and standards
- Desire for **mechanisms for SMEs to partner with other organisations or academia** to bolster limited internal expertise and resources**.**

GENERAL FINDINGS: INTERVENTIONS

3.

Communication across disciplines

- Desire for **common language and understanding** across disciplines
- Interest in assurance techniques that are **comprehensible to non-technical staff**
- Desire for **established definitions of foundational concepts** (e.g., “fairness” and “explainability”).

4.

Clear link between assurance and regulation

- Understanding **how AI assurance can support compliance with regulation** will be a **key motivator** for industry adoption
- Desire for **resources from regulators** that provide guidance that outlines **what is required to demonstrate compliance**
- Particularly important for **organisations that operate internationally**

HR & RECRUITMENT: OVERVIEW

In the HR and recruitment sector, AI systems are applied across a range of functions within the recruitment life cycle, including:

- Sourcing
- Screening
- Interview
- Selection

Adopting AI offers the automation and simplification of existing processes. However these technologies also pose novel risks. Specifically, these systems pose unique **rights-based harms** such as those that arise from **a lack of fairness.**

european actuarial academy

HR & RECRUITMENT: BARRIERS & INTERVENTIONS

Barrier	Intervention
Lack of knowledge/skills: • Many HR and recruitment organisations procure AI-enabled tools from **third party providers**. • These organisations often have **limited in-house AI expertise,** and may assume that requisite checks have been performed by the supplier.	**Sector-specific guidance** • CDEI / REC "Data-driven tools in recruitment guidance"
Lack of demand: • In this sector there is a lack of both **internal and external demand for assurance** to evaluate these systems (i.e., from senior leaders and end-users/customers, respectively).	**Resources to demonstrate the value-add of assurance** • CDEI AI Assurance guide
Regulatory uncertainty: • There is **no dedicated regulator for HR & recruitment**. As such, there is limited dedicated regulatory resourcing to provide guidance to support the compliance of AI with future AI regulation.	**Additional regulatory clarity** • HMG AI Regulation White Paper • Government collaboration with regulators to support implementation of the regulatory framework (ongoing)

FINANCE: OVERVIEW

AI and data-driven approaches are being applied across a range of functions in the financial services sector, including:

- Fraud detection and anti-money laundering
- Customer interactions
- Risk management
- Compliance

Adopting AI offers a wealth of potential benefits. However these technologies also pose novel risks. Specifically, these systems pose unique **rights and privacy-based harms** such as those that arise from **a lack of fairness and transparency.**

european actuarial academy

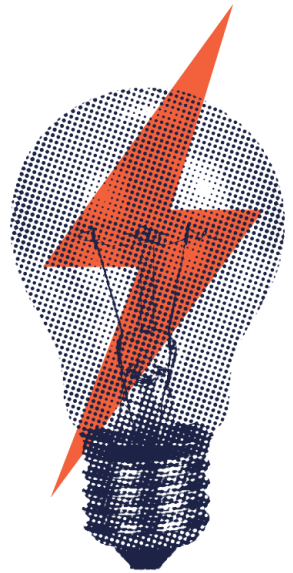
FINANCE: BARRIERS & INTERVENTIONS

Barrier	Intervention
Lack of knowledge/skills: • The number of people in the finance sector with relevant AI assurance training or skills remains relatively low • A need for **upskilling staff on concepts around assurance, ethical AI, and responsible innovation.**	**L&D to increase awareness of AI assurance** • CDEI / ATI "Introduction to AI assurance" e-learning module
Lack of awareness of available assurance techniques: • Many organisations **don't know what assurance techniques exist**, or which technique(s) should be used to evaluate a particular system.	**Repository of AI assurance techniques** • OECD Catalogue of tools and metrics for Trustworthy AI • CDEI Portfolio of AI assurance techniques (forthcoming)
Lack of signposted best practice: • Many finance organisations have governance processes in place to support non-AI related assurance practices - most commonly, financial audit. • However, participants reported **lack of clarity around how to adapt existing governance frameworks to address novel AI-related risks.**	**Signposted best practice** • HMG AI Regulation White Paper • Government collaboration with regulators to support implementation of the regulatory framework (ongoing)

CONNECTED AND AUTOMATED VEHICLES: OVERVIEW

CAV typically involve multiple algorithms, each developed and designed for a specific purpose, including:

- Object identification and classification
- Object localisation
- Route planning and optimisation
- Automated decision-making



CAV offer a wealth of benefits, from minimising errors in human driving, to reducing congestion and pollution. However, the adaptive and autonomous nature of these systems may pose risks **to health, safety and security of the vehicle and road users**.

european actuarial academy

CAV: BARRIERS & INTERVENTIONS

Barrier	Intervention
Lack of awareness of available standards: • Many organisations are **not aware of the existence of CAV-related standards** • This was reflected in respondents' selection of "relevant standards not available" as a barrier to adoption, despite the publication of many standards in this domain.	**Resources to identify/select standards** • UK AI Standards Hub
Lack of mechanisms to recognise assurance efforts: • There very **few governance mechanisms** (i.e., certification, kite marking) **to recognise whether CAV manufacturers have adopted AI assurance techniques/standards.** • Desire for such mechanisms to **demonstrate compliance to customers.**	**Mature governance and regulatory landscape** • CDEI/CCAV "Responsible innovation in self-driving vehicles" report
Lack of signposted best practice: • There is **limited guidance** that advises on which AI assurance techniques to use and when. • Need for tools to **aid the selection and application of assurance techniques** at each stage of the AI lifecycle.	**Examples of good practice** • CDEI Portfolio of AI Assurance techniques (forthcoming)

SUMMARY AND NEXT STEPS

Barrier type	Barrier to using AI assurance	Potential intervention
Workforce barriers	Lack of knowledge/skills	General L&D & sector-specific guidance • CDEI / ATI e-learning module on "Introduction to AI assurance" • CDEI / REC "Data-driven tools in recruitment guidance"
	Lack of awareness of available assurance techniques	Toolkit of AI assurance techniques • OECD Catalogue of Tools and Metrics for Trustworthy AI
	Lack of awareness of technical standards	AI standards repository • AI Standards Hub
Operational/ market barriers	Lack of demand	Demonstrate value add of assurance • CDEI AI Assurance guide
	Lack of mechanisms to recognise assurance efforts	Mature governance & regulatory landscape • CDEI/CCAV "Responsible innovation in self-driving vehicles" report
	Lack of signposted good practice	Examples of good practice • CDEI Portfolio of AI Assurance techniques
Governance barriers	Regulatory uncertainty	Additional regulatory clarity • HMG AI Regulation White Paper • Government collaboration with regulators to support the implementation of the forthcoming regulatory framework (ongoing)

6.RESOURCES AND NEXT STEPS

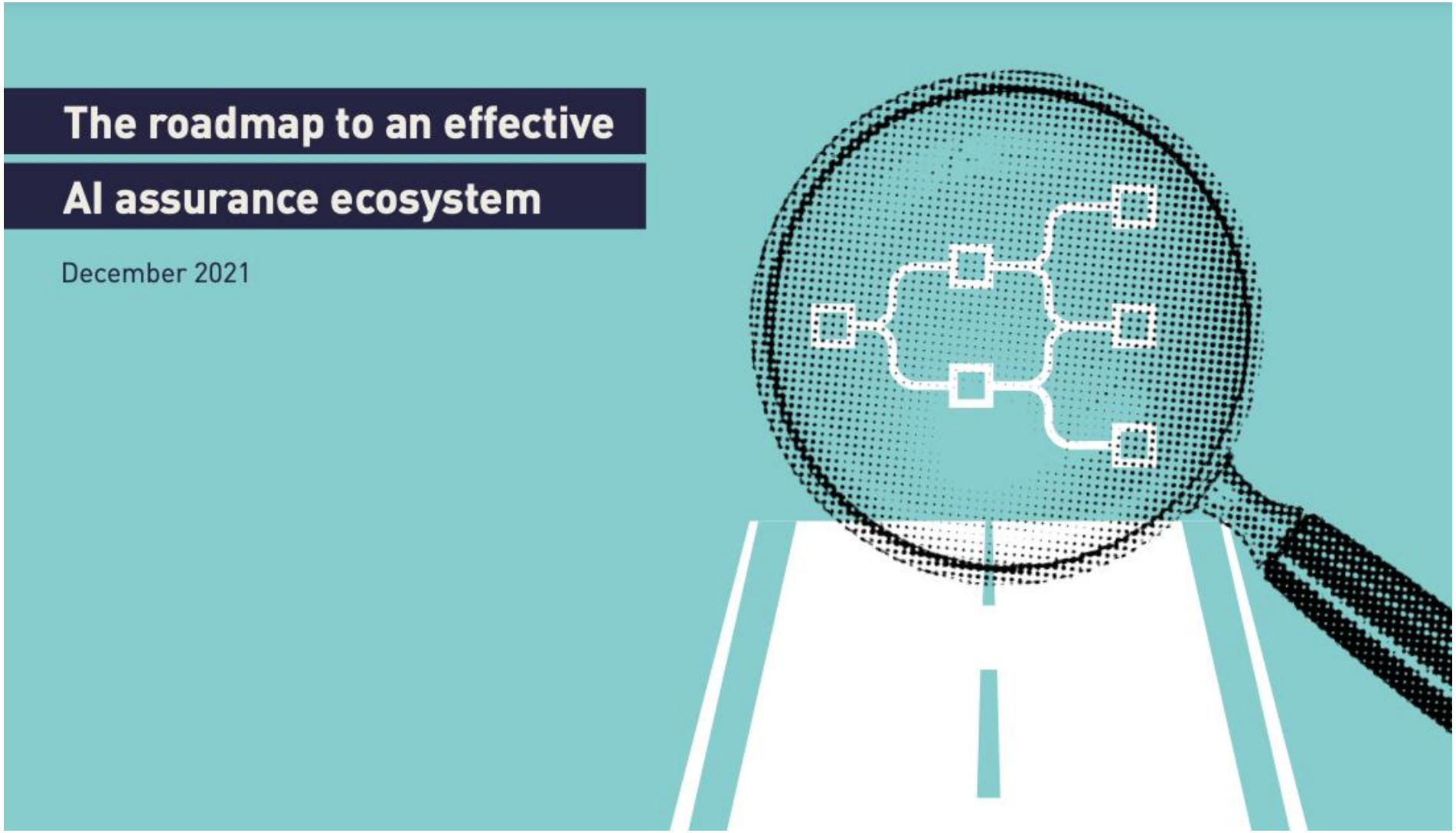
NEXT STEPS

CDEI workstreams to promote the development and adoption of tools for trustworthy AI

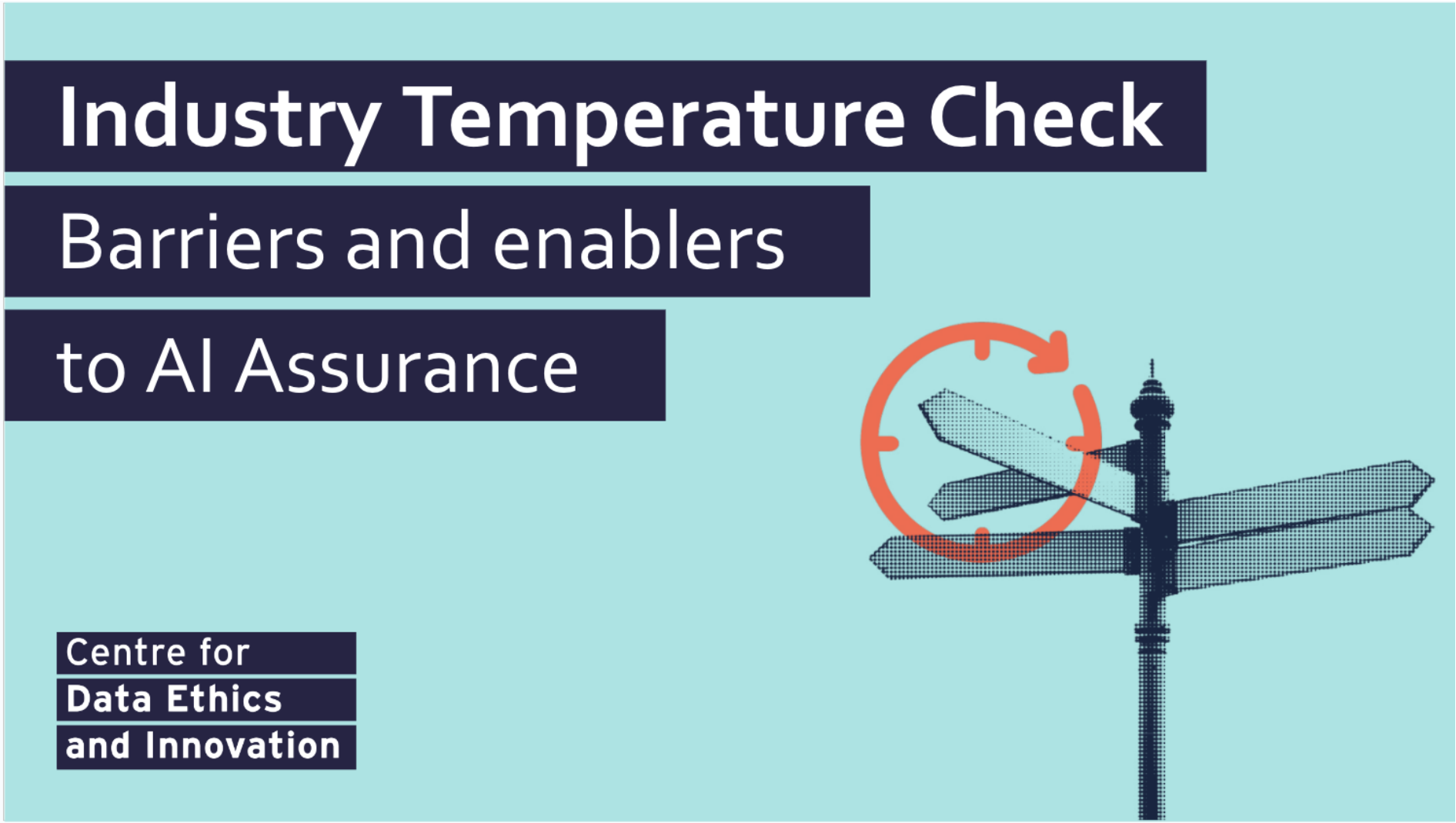
Portfolio of AI Assurance Techniques

Fairness Innovation Challenge

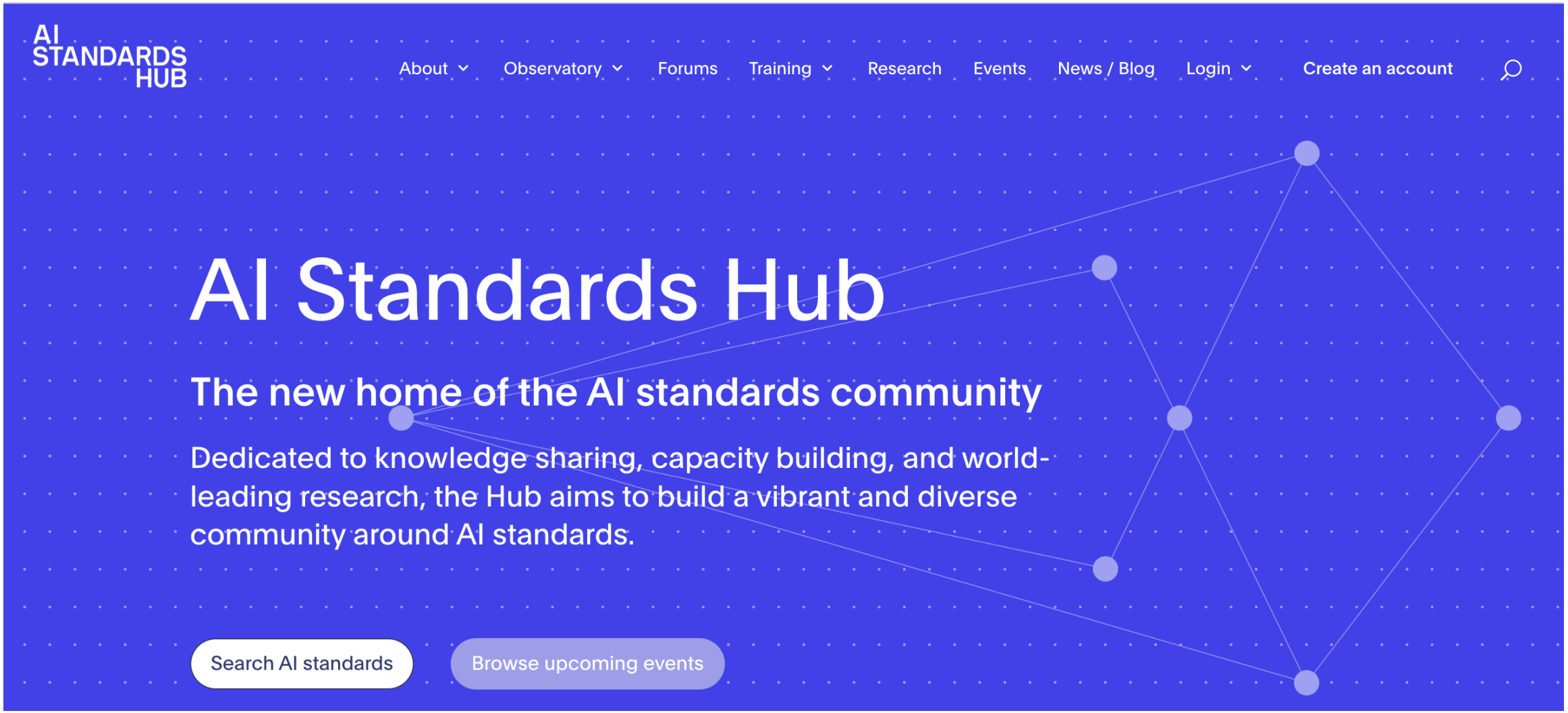
Stakeholder workshops



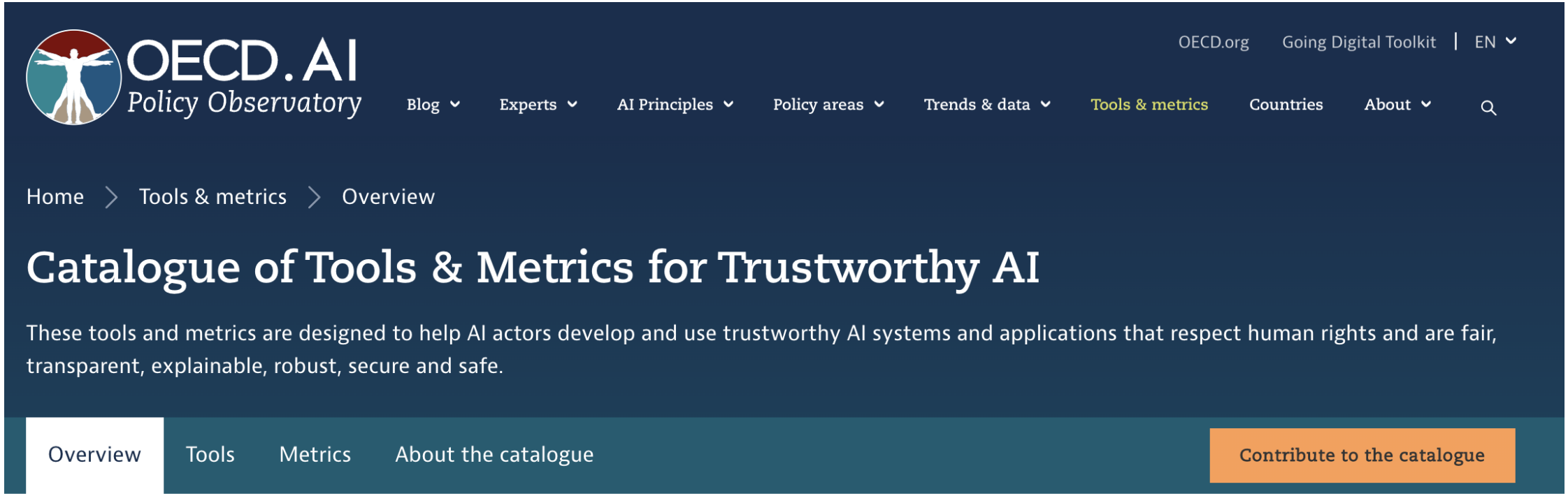
https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem



https://www.gov.uk/government/publications/industry-temperature-check-barriers-and-enablers-to-ai-assurance



https://aistandardshub.org/



https://oecd.ai/en/catalogue/overview

Nuala Polo is a Senior Policy Advisor at the Centre for Data Ethics and Innovation (CDEI) working in the AI Assurance team. Her work focuses on exploring the role of tools for trustworthy AI, like assurance techniques and standards, to manage risks, build trust, and support AI governance. She was the lead author of the CDEI's Industry Temperature Check report, which identified industry barriers to engaging with tools for trustworthy AI in the HR & recruitment, finance, and connected and automated vehicle (CAV) sectors.

Nuala holds an MSc Cognition in Science and Society from the University of Edinburgh, with a focus on AI Ethics, and BA Combined Honours in History of Science & Technology and Philosophy with a minor in Mathematics from the University of King's College. Prior to her role at CDEI, Nuala worked as a consultant for an AI Ethics consultancy firm, and led dissemination and communication activities for Horizon-2020 funded research projects focused on the ethical implications of disruptive and emerging technologies, including SHERPA and TECHETHOS.

ABOUT ME

Nuala

Polo

Centre for
Data Ethics
and Innovation

Thank you very much for your attention

EAA e-Conference on
Data Science & Data Ethics

16 May 2023

Contact

Nuala Polo

Senior Policy Advisor, CDEI

nuala.polo@dcms.gov.uk