TRILATERAL RESEARCH
Ethical AI

Innovate UK

BridgeAI

Executive Summary of the Report on the Core Principles and Opportunities for Responsible and Trustworthy AI

Ref: PS22477

Contents

1. Purpose

The UK Science and Technology Framework, published in March 2023, identifies AI as one of five critical technologies in which the UK is seeking to build a strategic and globally competitive advantage in order to become a science and technology superpower by 2030. According to a Capital Economics report commissioned by the Department for Digital Culture, Media and Sport (DCMS), expenditure on AI is expected to grow to around £30bn in 2025 and to £83.5bn by 2040 at a compound annual growth rate of 8.4%. This growth will lead to an acceleration of the uptake of AI systems by businesses as well as consumers, creating a positive feedback loop for the overall AI market. By 2030, AI will be a major driver for businesses and will transform every sector of the economy.

The potential for commercial growth in this sector reflects the current widespread use of AI systems throughout society. This pervasive use of AI is likely to continue to have a significant impact on individual wellbeing, democratic values, fundamental rights and a sustainable environment. Therefore, a key aspect of the UK strategic vision to promote the long-term social and economic benefits of AI, and to leverage the innovation and market potential of AI, is the creation of an ecosystem for AI systems to be responsible and trustworthy.

This report represents a single and common frame of reference on core principles, key innovation priorities, new commercial opportunities and policy and standards development relating to responsible and trustworthy AI (RTAI) for the UK. It creates a shared language to more easily communicate commercial innovation opportunities to stakeholders in the industry. It also establishes a framework of RTAI focused on maximising societal benefits and protecting fundamental rights. This report identifies and evaluates key steps for the UK to lead in RTAI by providing a prioritisation of opportunities to inform future investments in research, innovation and policy/standards development to achieve the economic and societal benefits from RTAI in the long term. It concludes that the clearest opportunities for innovation, market capture and policy/standards development can be found in AI assurance, sustainable AI and the sociotechnical development of AI systems.

2. Core Findings

Section 2 establishes the core principles of RTAI throughout the AI lifecycle in order to formulate the UK vision of RTAI and to derive key innovation priorities, commercial opportunities and avenues for policy/standard development. It includes guidance on how to operationalise the core RTAI principles as well as challenges to achieving this goal. The principles were distilled from an analysis of over 40 academic and international RTAI policy documents including well-established UK guidance documents. The outcome of this section is the establishment of a unified vison and concrete foundation for RTAI as a necessary step in advancing the status of the UK as a global leader in responsible and trustworthy AI.

The key findings of this section are:

- Core principles to develop, deploy and use responsible and trustworthy AI include: (1) appropriate transparency and explainability; (2) safety, security and robustness; (3) non-maleficence; (4) fairness and justice; (5) privacy and data protection; (6) accountability.

- Agreement between UK principles and those accepted internationally ensures that the principles adopted by the UK can be exported, thereby facilitating commercial trade and the dissemination of UK thought leadership in the field.

- These principles need to be operationalised through a series of tools, processes, standards, policies and regulations.

- Efforts to sustain or promote the trustworthiness of an AI system should be directed at all stages of the AI lifecycle.

- Realisation of an ecosystem for trustworthy and responsible AI requires efforts and expertise of all AI stakeholders.

Section 3 identifies key innovation priorities, derived from the RTAI principles, that UK funding bodies can use as a source of evidence to promote responsible and trustworthy innovation. To ensure the achievement of the economic and societal benefits from RTAI in the long term, the UK can drive the implementation and further development of the RTAI principles through research and innovation funding. Given the rapid pace of innovation in AI, RTAI innovations need to be valid across contexts and technologies and continue to be aligned

with the advancement of new and emerging trends in AI. The information presented in this section demonstrates that there is a significant opportunity for the UK to be a leader in propelling the development of innovative sociotechnical and environmentally sustainable methods and processes.

The benefit for stakeholders of this section of the report is identifying resources for AI developers to support innovative RTAI development. For researchers and research funding bodies, the benefit is identifying areas to explore and support with research investment, and for policymakers to understand the areas that may need policy support to encourage investment in those areas that are fundamental to achieving trustworthy AI.

The key findings of this section are:

- Approaches to assessing the potential social, ethical and legal impacts of AI systems, and how they impact upon core principles exist. However, to the extent that these are voluntary, the consistency and quality of implementation suffers, and they do not yet contribute sufficiently to AI assurance or ensure that the core principles are met.
- High-level guidance setting out broad approaches to RTAI also exist. What is needed is more granular knowledge about appropriate implementation of these approaches in particular domains, industries, and contexts, including solving problems in those domains. This is particularly true if AI in the UK is to be regulated on a sectoral basis (see section 5).
- There is an emerging landscape of technological tools to support elements of RTAI, but tools are unevenly distributed across the core principles, leaving some under-supported.
- Significant innovation opportunities exist in methods and processes needed to operationalise RTAI. Investment in research and development in these areas can promote the status of RTAI domestically in the UK as well as advance the position of the UK as a global leader in RTAI. There are priority areas, especially in AI assurance and sustainable AI.

Section 4 provides a guide to industry stakeholders to help identify new avenues of development and growth in the UK deriving from RTAI. While AI capabilities, products, and services are widely advertised by companies, uptake of Responsible and Trustworthy AI (RTAI) is low. According to a 2022 report, only 6% of organisations have set the groundwork for RTAI, and 25% have yet to establish any meaningful RTAI capabilities. Another report from that year finds that a majority of companies have not taken key steps toward responsibility, such as reducing bias or ensuring that they can explain

AI-powered decisions.[1] Consulting firms have started to provide services for assuring that AI systems are responsible and transparent. As the risks of AI systems increase with their complexity and their adoption by businesses and society, there is tremendous potential for RTAI and for AI assurance services – both in assuring the AI systems of UK companies and in exporting assurance services and techniques abroad.

This section of the report identifies both commercial opportunities related to advancing RTAI principles within the AI market generally, as well as significant opportunities residing in the still latent AI assurance marketplace. Capitalising on these opportunities will help achieve both economic and societal benefits in the UK in the long term and drive UK market capture the RTAI sector internationally.

The key findings of this section are:

- Developers should implement RTAI principles in existing and new AI systems to increase revenue and public trust.
- Industry actors should exploit commercial opportunities where RTAI is underused. There are significant market opportunities in quantum computing and synthetic data.
- Industry actors should develop and market tools and services that facilitate RTAI and RTAI assurance.
- Regulators should address AI assurance market fragmentation.
- Policymakers should support the export of AI assurance.

Section 5 assesses current RTAI policies and standards in the UK and evaluates their level of maturity and appplication. Supportive regulation through policies and standards can demonstrate the UK's position as a global leader in RTAI. This section provides policy approaches from other countries and international organisations to enhance alignment in international cooperation and trade, and to identify where the UK can influence international activities in RTAI. To show where the UK can continue to advance its domestic RTAI objectives, this section gives an overview of existing standards and ongoing standardisation efforts. The main findings from this section demonstrate that the UK can drive improvements in policies and standards supporting: (1) AI assurance in order to foster the development of a consistent and coherent assurance ecosystem; and (2) sustainable AI to protect the environment and become a global leader in an urgent area of concern. These policy and standards initiatives will power

[1] See also TechUK, "AI Adoption in the UK: Putting AI into Action".

the realisation of societal and economic benefits of AI in the long term, while maximising wellbeing and protecting fundamental rights in the UK and beyond.

The key findings of this section are:

- Standards bodies to agree and develop measurement metrics for compliance. Technical standards for AI assurance techniques and services are seen as the most important tools for compliance with the responsibility and trustworthiness requirements for AI.
- UK government to devise and implement policies for sustainable AI, including considerations regarding the environmental footprint of different AI systems.
- UK government to ensure regulators are sufficiently empowered and adequately resourced to implement the proposed AI regulatory framework. This could involve bringing forward plans to place the AI principles on a statutory footing, as well as clarifying and making provision for the additional funding that regulators may require, particularly cross-sector regulators such as the ICO.
- UK government to clarify, through the proposed AI Regulation Roadmap, and implement the range of central support functions designed to support overall coordination of the AI regulatory framework.
- UK government to strengthen the proposed AI regulatory framework by creating a responsibility and liability framework for demonstrating compliance with AI regulatory principles, applicable to all AI lifecycle actors.

3. Impact

Policymakers, funding bodies, industry stakeholders and standards bodies can use the information and guidance in the report to continue to advance the UK's position as a global leader in RTAI.
There is no question that AI systems will continue to impact almost every sector of society and affect individual lives in significant ways. Even though technological development progresses rapidly and in sometimes unforeseeable directions, the opportunities to steer these systems to be responsible and trustworthy are already clear. Promoting innovations, capitalising on commercial opportunities and establishing regulatory consistency focused on AI assurance, sustainable AI and sociotechnical methods will solidify the UK as forerunner in political, scientific, technological and commercial sectors and leverage the social and economic benefits of RTAI in the long term.


TRILATERAL
RESEARCH
Ethical AI


UKRI
Innovate
UK


BridgeAI