TRILATERAL RESEARCH
Ethical AI

UKRI Innovate UK

BridgeAI

Report on the Core Principles and Opportunities for Responsible and Trustworthy AI

Ref: PS22477

Abbreviations

AI	Artificial Intelligence
AIPPF	AI Public Private Forum
BoE	Bank of England
BSI	British Standards Institute
CAV	Connected and autonomous vehicle
CDDO	UK Central Digital and Data Office
CDEI	Centre for Data Ethics and Innovation
CEN	European Committee for Standardization
CEN-CLC/JTC 21	The CEN/CENELEC Joint Technical Committee 21 on Artificial Intelligence
CENELEC	European Committee for Electrotechnical Standardization
CoE	Council of Europe
DCMS	UK Department for Digital Culture, Media and Sport
DRCF	Digital Regulation Cooperation Forum
DSIT	UK Department for Science, Innovation and Technology
EbD	Ethics-by-Design
ESG	Environmental, social and governance domain
ETSI	European Telecommunications Standards Institute
FCA	Financial Conduct Authority
GDP	Gross domestic product
GenAI	Generative Artificial Intelligence
HEI	Higher Education Institution
HR	Human Resources
HUDERIA	Human Rights, Democracy, Rule of Law Impact Assessment
HVAC	Heating, ventilation and air conditioning
IEC	International Electrotechnical Commission
ICO	UK Information Commissioner’s Office
ISO	International Organization for Standardization
LLM	Large language model
ML	Machine learning
MLOps	Machine learning operations
NGO	Non-government organisation

NHS	National Health Services (UK)
NIST	National Institute of Standards and Technology (USA)
NLP	Natural language processing
OECD	Organisation for Economic Co-operation and Development
RTAI	Responsible and trustworthy artificial intelligence
SDO	Standard development organisation
SSH	Social sciences and humanities
SME	Small and medium-sized enterprise
STEM	Science, technology, engineering and mathematics
UKRI	UK Research and Innovation
UNESCO	United Nations Educational, Scientific and Cultural Organisation

Contents

1. **Introduction** 5
    - Purpose of this report 5
    - Relevant audiences 5
    - Structure 5
2. **Core principles of RTAI** 7
    - UK RTAI principles 8
    - Principles globally 13
    - Challenges to application 14
    - Application of the core principles 16
    - Benefit of this section to stakeholders 18
3. **Key innovation priorities** 19
    - Existing guidance documents and tools for achieving RTAI 19
    - Barriers to RTAI 25
    - Innovative methods needed to achieve RTAI 27
    - Benefit of this section to stakeholders 30
4. **Commercial opportunities** 31
    - Commercial opportunities for RTAI principles 32
    - Global market leadership in assurance 41
    - Benefit of this report to stakeholders 43
5. **Policy and standards analysis** 44
    - International policies on responsible and trustworthy AI 45
    - UK standardisation efforts in responsible and trustworthy AI 47
    - Scoring UK policy and standards 49
    - Benefit of this report to stakeholders 58
6. **Conclusion** 60

Annex I: Innovation opportunities in responsible and trustworthy AI 61

Annex II: Current standards relevant to responsible and trustworthy AI 66

References 77

1. Introduction

PURPOSE OF THIS REPORT

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.

RELEVANT AUDIENCES

This report was prepared for Innovate UK, a part of UK Research and Innovation (UKRI). Additionally, the widespread use of AI systems in society reveals their potential impact on individual wellbeing, democratic values, fundamental rights and a sustainable environment. Owing to this potential impact, responsible and trustworthy AI affords crucial innovation and commercial opportunities. Consequently, the findings in this report are also relevant for policymakers, standards bodies, regulators, research funding organisations, industry stakeholders and AI developers.

STRUCTURE

Section 2 describes the framework of core principles underlying RTAI for the UK. It outlines high-level requirements to apply these principles to AI systems. The framework establishes the foundation from which to derive opportunities in RTAI innovation, industry and policy/standards. This section identifies challenges in applying these core principles as well as case studies of untrustworthy and irresponsible AI. These case studies highlight the complexity associated with implementing the principles and the need for the development of innovative methods and commercial investment in RTAI.

Section 3 identifies innovation gaps that, if filled, would significantly promote the development of RTAI in the UK. 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. Since AI systems are tools made by and for people, there is a clear opportunity for innovation in methods that combine both social and technical aspects of all stages of the AI lifecycle.

Section 4 provides a guide to industry stakeholders to help identify new avenues of development and growth in the UK deriving from RTAI. Since RTAI core principles apply to every AI system, the commercial opportunities are far-reaching, including numerous opportunities in the AI assurance marketplace and in creating environmentally sustainable AI.

Section 5 provides opportunities for the UK to influence international RTAI policies as well as a "scoring" assessment of the maturity of polices and standards that reflect the core RTAI principles. The outcome of this assessment reveals where the UK is already leading in relevant policies and standards and where it can continue to advance its position as a global leader in RTAI. Aligned with the findings in the preceding sections, there is opportunity for further advancement in AI assurance, sustainable AI and the sociotechnical development of AI systems.

Annexes I & II contain a comprehensive list of innovation opportunities in RTAI and an exhaustive list of UK standards relevant to RTAI, respectively.

2. Core principles of RTAI

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. In order to achieve this goal, and to fully leverage the innovation and market potential of AI, AI systems must be responsible and trustworthy.

To be **trustworthy**, an AI system must be lawful, robust and ethical, making it worthy of the trust of both end-users as well as other affected parties. **End-users**, which can include individuals or organisations deploying and using these systems, must be able to have a satisfactory level of confidence that the system is achieving its intended objectives while protecting their interests and/or rights. They must also be confident that systems are not infringing on the rights and interests of others. Similarly, **affected parties**, which can include individuals, organisations or the broader public, should be able to trust that AI systems that impact their lives will function without introducing harmful or unfair risks to them or the environment.

Responsible AI requires actors across the AI system's lifecycle to acknowledge and act upon their duties to protect the interests and rights of end-users and affected parties. Through the development, deployment, and use of AI, different people, from CEOs to AI developers to end-users, make key decisions about the functioning of the systems that impact the interests and rights of others and the sustainability of the environment. These individuals are accountable for their decisions and actions.

This section establishes the **core principles of RTAI** throughout the AI lifecycle in order to derive key innovation priorities, commercial opportunities and avenues for policy/standard development and to position the UK as a global leader in responsible and trustworthy AI. First, this section identifies the core principles that will formulate the UK vision of RTAI. Next, it compares these principles with RTAI principles promoted by other countries. 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. However, establishing a framework of core principles is insufficient on its own to position the UK as a global leader in RTAI; these principles must also be operationalised. This section ends with guidance on how to operationalise the core RTAI principles as well as challenges to achieving this goal. They main takeaways from the section are:

- Efforts to sustain or promote the trustworthiness of AI systems should be directed at all stages of the AI lifecycle.
- The realisation of an ecosystem for trustworthy and responsible AI requires efforts and expertise of all AI stakeholders.

The information presented in this section is critical to establishing innovation priorities, identifying commercial opportunities and developing regulations and standards in RTAI, which are described in subsequent sections of the report.

UK RTAI PRINCIPLES

AI systems will likely continue to have a widespread and deep impact on individuals, society and the environment. As such, guiding principles throughout the AI lifecycle are needed for the development, deployment, and use of AI systems. Furthermore, establishing a framework of these core principles will derive key innovation priorities, commercial opportunities and policy needs for UK policymakers, funding bodies and industry actors. These priorities and opportunities provide concrete, actionable steps for these stakeholders to leverage the long-term social and economic benefits of AI and advance the UK's position as a global leader in RTAI.

The following list of core RTAI principles are the result of an analysis of over 40 academic and international RTAI policy documents including well-established UK GOV guidance documents such as 'A pro-innovation approach to AI regulation' and Understanding artificial intelligence ethics and safety from The Alan Turing Institute, as well as from influential international bodies such as the EU High-Level Expert Group on Artificial Intelligence and the Organisation for Economic Co-operation and Development (OECD). The principles in the framework will serve to:

Maximise societal wellbeing and protect fundamental rights,

Elicit key innovation priorities and commercial activities (sections 3, 4),

Identify avenues for policy/standards development (section 5)

Ensure international alignment and interoperability to promote global cooperation, trade and UK thought leadership internationally.

The framework consists of six high-level RTAI principles:

Appropriate transparency and explainability refers to duties to **document** and allow access to relevant information about AI systems and to present this information in a way that is **comprehensible** to stakeholders including developers, end-users, and affected parties. The information ought to include how and why an AI system produced its output, its intended purposes, its limitations and capabilities and its projected benefits and risks. Satisfying the principle of transparency and explainability is **a prerequisite to implementing the entire RTAI framework** as it facilitates compliance with the other principles and promotes the trustworthiness of, and responsibility for, the AI system. Documenting and explaining the functioning, scope and objectives of an AI system augments traceability and auditability as well and provides information to stakeholders to make informed decisions.

WHY?

Explainability is difficult to ensure for machine learning technology, and in particular for deep learning systems using neural networks. These systems produce results in ways that are difficult to decipher, even for highly trained AI developers. For this reason, these systems are said to function like “black boxes”.

Safety, security and robustness are essential RTAI components as the performance of AI systems must be robust enough to function as intended, both in testing stages and in real-world settings. If an AI system does not function as intended, the system could cause harm. A robust system is accurate, and its outputs are reproducible and reliable. **Safety** requires the prevention of harm caused by a malfunction of the system. **Security** requires the implementation of appropriate cybersecurity measures to protect against hacking or other adversarial attacks.

WHY?

Systems based on machine learning are subject to threats such as data poisoning, which involves tampering with training data to produce undesirable outcomes.

Non-maleficence requires AI systems to do no harm against individuals, communities, society at large and the environment. Harm includes violations of **human dignity** and **human rights** as well as of mental, physical, and **environmental integrity** and **well-being**. Although this principle has some similarities to the previous one, it differs because even a safe, secure, and robust system can cause harm: for example, when a functioning system is misapplied to pose a threat to people.

WHY?

Large Language Models (LLMs) have large carbon footprints as they require a vast amount of energy to develop and to operate. This impact could harm the environment and worsen climate change.

Privacy refers to the right to limit access to, and augment a person's control over, personal information. **Data protection** and **cybersecurity** regulations and best practices help protect the right to privacy. Strongly related to privacy are concerns about the mass collection of digital information and the potential for clandestine or blanket surveillance facilitated by some AI systems.

WHY?

Facial recognition software used for surveillance purposes in retail stores have drawn privacy concerns from civil society representatives and society at large.

Fairness & justice refer to duties to promote equality, equity and non-discrimination. Fairness is crucial due to the **harmful biases** AI systems have the potential to perpetuate. Biases in AI can lead to entrenching structural social inequalities and stereotypes affecting people in vulnerable situations. The principle of fairness and justice also requires **fair compensation** for the work required to build, train and maintain AI systems, including labelling datasets or flagging content. Furthermore, existing inequalities between privileged and less privileged groups, as well as developed and developing countries, can be exacerbated by poorly accessible or low-quality internet, scarce digital services or a shortage of resources to enhance digital skills and digital literacy. Finally, AI systems must be developed in **compliance** with existing **laws, human rights, and democratic values**.

WHY?

An algorithm can have an adverse effect on vulnerable populations even without explicitly including protected characteristics. This often occurs when a model includes features, called proxies, that are correlated with these characteristics.

Accountability requires individuals or organisations to take ownership of their actions or conduct and to explain reasons for which decisions and actions were taken. When mistakes or errors are made, it also implies establishing accessible avenues for **contestability** and **redress** and taking action to ensure a better outcome in the future (for example, by retraining the model). Accountable organisations or individuals will ensure the proper functioning of AI systems throughout the AI lifecycle. They are expected to do so in accordance with their roles and applicable regulatory frameworks, and to demonstrate this through their actions and decision-making processes. Accountability requires continual human understanding of an algorithm and its output. Hybrid decision-making and human-in-the-loop approaches introduce human control by allowing humans to interact with algorithmic output at every stage of the AI lifecycle. Finally, to ensure ongoing governance of AI systems, companies that develop and deploy AI ought to establish a shared responsibility model with end-users.

WHY?

Self-driving cars can cause accidents. To what extent either the car maker or the driver of the car is liable for such an accident, and under what conditions, requires careful consideration.

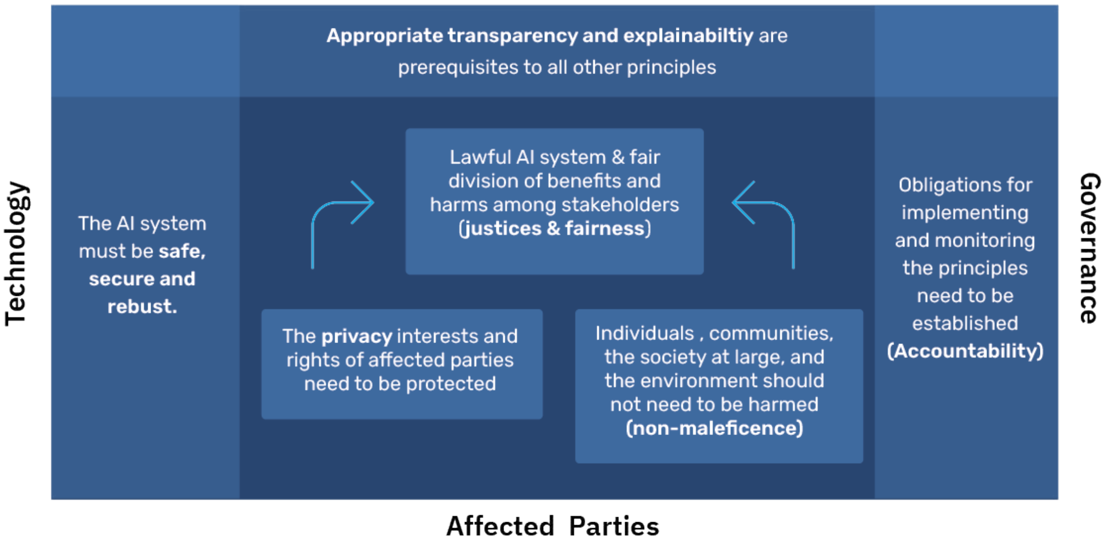
Appropriate transparency and explainabiltiy are prerequisites to all other principles
Technology
The AI system must be **safe, secure and rebust.**
Lawful AI system & fair division of benefits and harms among stakeholders **(justices & fairness)**
The **privacy** interests and rights of affected parties need to be protected
Individuals , communities, the society at large, and the environment should not need to be harmed **(non-maleficence)**
Obligations for implementing and monitoring the principles need to be established **(Accountability)**
Governance
Affected Parties

Figure 1: The six RTAI principles and their interaction.

Figure 1 shows how the principles relate to each other. The principle 'appropriate transparency and explainability' is a prerequisite to all other principles. The principle of 'safety, security, and robustness' is a technological requirement. The principles of 'privacy', 'non-maleficence', 'justice and fairness' have direct impact on affected parties. The principle of 'accountability' has a strong governance component to it and is associated with the obligation to implement and monitor the principles and respond to the needs of affected parties.

These core principles present **high-level guidance** for the development, deployment, and use of RTAI throughout the AI lifecycle. They form a common language and point of reference to derive key innovation priorities (section 3), commercial opportunities (section 4) and avenues to develop RTAI-relevant policies and standards (section 5). In combination, these subsequent sections will identify long-term social and economic benefits of AI and provide actionable guidance to position the UK as a global leader in RTAI.

PRINCIPLES GLOBALLY

An international consensus[1] on the core principles of RTAI will allow for **international alignment and interoperability**. Table 1 makes a comparison of RTAI international guidance documents based on the central documents from each country listed. The table shows a high level of international agreement with the core RTAI principles presented in this report, which provides an opportunity for the UK to drive international alignment in RTAI.

RTAI Principles	UK 2023	EC 2019	OECD 2019	UNESCO 2021	IEEE 2019	US 2022	China 2021	India 2021	Japan 2019	Uruguay 2019	Russia 2021
Appropriate transparency & explainability	●	●	●	●	●	●	●	●	●	●	
Safey, security and rebustness	●	●	●	●	●	●	●	●	●	●	
Non-malficence	●	●	●	●	●	●	●		●	●	●
Fairness and justice	●	●	●	●	●	●	●	●	●	●	●
Privacy and data protection	●	●	●	●	●	●	●	●	●	●	
Accountability	●	●	●	●	●		●	●	●	●	

Table 1: An overview of the principles in international frameworks.

[1] Nonetheless within this body of international frameworks, some regions appear underrepresented in the international discussion on RTAI, including Africa, Latin America and the Caribbean, and Central Asia. This lack of representation requires monitoring to ensure acceptance of these principles in these areas in the future.

Table 1 shows strong international consensus on each of the UK's core RTAI principles. This general agreement on principles provides the UK with an opportunity to influence international RTAI policies, advocate for international adoption of its own RTAI framework and capitalise on international export and trade in RTAI products.

However, the application of the core principles has faced challenges resulting in irresponsible and untrustworthy AI. Identifying these challenges in the next subsection can help stakeholders avoid future mistakes as they seek to exploit the opportunities set out in sections 3, 4 and 5.

CHALLENGES TO APPLICATION

At **every stage of an AI system's lifecycle**, the responsibility and trustworthiness of the system can be strengthened or weakened. AI systems can be transformative and have unforeseen impacts making it challenging to accurately identify these impacts in advance. For example, at the final stages of its lifecycle, the environmental costs of disposing the material infrastructure supporting AI could be unfairly placed on countries and regions with environmental laws offering lower protections.

Furthermore, RTAI principles can sometimes **conflict with each other**. In such cases, it is necessary to make **trade-offs**. For instance, transparency requirements can affect the robustness, safety, and security of a system because information about how the system works provides the necessary information to launch adversarial attacks. As trade-offs emerge, they should be carefully monitored for risks to the core principles. Additionally, these trade-offs should be documented and managed under the principles of transparency and accountability. If **no acceptable trade-offs** exist – such as when the AI system violates fundamental rights or is disproportionately untrustworthy or irresponsible – then an AI system should not be developed nor used.

These concerns are not merely hypothetical. Three case studies illustrate untrustworthy and irresponsible AI:

- Several scandals have occurred around government use of predictive algorithms for fraud prevention in welfare, including the Child Benefits scandal in the Netherlands, the 'Robodebt' scandal in Australia and the National Institute of Social Security in Spain. Whilst the systems vary in complexity, they have all had issues with predictive accuracy, reliability, accountability, a lack of transparency, an ability to contest decisions made by these systems or gain redress. These issues required addressing by the AI developers and end-users in each stage of the AI lifecycle. The system in the Netherlands was also found to be discriminatory. These systems have resulted in significant impact on people wrongly accused of fraud, with some victims dying by suicide. In the Netherlands case, the scandal resulted in the resignation of the entire government, and in Australia $2bn AUS was repaid to victims.

- Clearview AI Inc, a US-based company, scraped over 20bn images of people from the open web and social media to build a facial recognition service. This issue should have been stopped in the data collection stage of the AI lifecycle. The Information Commissioner, as well as regulators in Italy, Greece and France found significant breaches of data protection law. The ICO concluded that the processing of personal data to build and operate the service was not fair, lawful or transparent. The company was fined £7.5m and ordered to stop obtaining and using the data of UK residents. The EU regulators reached similar findings, and each levied the maximum €20m fines for breaching the GDPR. The European Parliament's recent version of the AI Act seeks to prohibit this use of AI.
- Computer vision technologies lacking reliability in semi-autonomous vehicles can result in dangerous driving. The US vehicle safety regulator, the NHTSA, opened an investigation into Tesla's 'Full Self-Driving Beta' system, leading to a recall notice for 360,000 cars, and requiring a software update to correct issues with unsafe behaviour around junctions, stop signs, and insufficient response to driver input. This case demonstrates the complexity of full automation in complex environments, and the consequences for unpredictable failures. Developers should have adequately addressed these issues in the development and testing stages of the AI lifecycle.

These examples illustrate that despite widespread agreement on the core principles of RTAI, serious transgressions of the principles continue. In order to protect fundamental rights and promote societal wellbeing as well as to leverage long-term economic benefits, there is a responsibility among all stakeholders to implement the RTAI principles throughout **all stages of the AI lifecycle**. The stakeholders and their responsibilities are identified in the next subsection.

APPLICATION OF THE CORE PRINCIPLES

Assessing the challenges arising from the application of the core principles demonstrates that the operationalisation of RTAI principles requires the efforts and expertise of all stakeholders (see Figure 2). Consequently, all stakeholders should identify their responsibilities to implement the RTAI principles both in terms of short- and long-term activities.

1 **Developers** and **industry** representatives can help to ensure an AI system is designed and developed in compliance with the core principles of RTAI, for instance by implementing security-by-design approaches to protect the safety, security and robustness of the system.

2 **Government bodies, regulators, and other policymakers** can help create the conditions to advance the core principles, monitor compliance and penalise their violations.

3 **Standards bodies** can define and help implement technical standards for best practices for RTAI systems. The effect of standardised best practices is to decrease competing interpretations and operationalisation of RTAI principles along the AI lifecycle.

4 **Funders** can fund projects and initiatives that explicitly seek to advance the development of RTAI.

5 **Shareholders and investors** can help create financial incentives to guide RTAI and choose to invest in RTAI even at an increase in price. In doing so, they commit to financially supporting the environmental, social, and governance (ESG) domain.

6 **Users and customers** can recognise and prioritise responsibility and sustainability alongside cost when evaluating AI systems and services or refrain from using untrustworthy and irresponsible AI systems.

7 **Researchers**, including from the fields of science, technology, engineering, and mathematics (STEM), and social science and humanities (SSH), can contribute to the political, economic, and public debate with scientific insights, and by discovering innovations to promote RTAI principles.

8 **Advisory bodies**, such as ethics and compliance bodies, can define, review and advise about best practices for RTAI systems.

9 The inclusion of insights from **representatives of the public** and **people in vulnerable situations** can inform the design, research and governance of the systems to ensure that standards, policy, and innovations resonate within society. Social actors, including civil society organisations, artists, and activists, play an active role in shaping public debate and the use and effects of technology.

BENEFIT OF THIS SECTION TO STAKEHOLDERS

Establishing a unified vison and concrete foundation for RTAI is a necessary step in advancing the status of the UK as a global leader in responsible and trustworthy AI. Comparing the UK core principles set out in this report with international principles demonstrates an agreement about their content and import, which can provide UK policymakers with confidence that the UK RTAI vision can be exported internationally, thereby enhancing international market capture and disseminating UK thought leadership. However, given the clear complexity raised by application of these principles, the subsequent sections will identify key innovation priorities, commercial opportunities and new avenues for policy and standards to achieve the economic and societal benefits from RTAI in the long term.

Key takeaways to create a common language and point of reference for RTAI:

- 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.
- 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 effort and expertise from all AI stakeholders.

3. Key innovation priorities

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. The framework of core RTAI principles (section 2) creates a common frame of reference and shared language to pursue this goal. This section identifies key innovation priorities, derived from those principles, that UK funding bodies can use as a source of evidence to promote responsible and trustworthy innovation. (Whilst this section identifies innovation priorities, a comprehensive list of innovation gaps can be found in Annex I.) This section begins by cataloguing existing guidance documents and current tools for implementing RTAI in order to document where progress in RTAI has already been made. The catalogue further functions as a shared reference guide for AI developers to implement RTAI. It also describes why barriers to RTAI development have occurred to explain where RTAI innovation is needed to avoid such barriers in the future. Combined, 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.

EXISTING GUIDANCE DOCUMENTS AND TOOLS FOR ACHIEVING RTAI

This subsection catalogues exiting procedures and technical tools for product development companies to implement the core principles of RTAI across the AI lifecycle. These state-of-the-art methods and tools are currently available for implementation by AI developers, however they also and reveal where current approaches end and RTAI innovation can begin.

There are two primary ways for AI product development companies to implement RTAI. First, they can develop and engage in **RTAI procedural measures** including ethics-by-design measures, such as impact assessments and traceability documentation. Second, they can implement **RTAI technical measures**, such as systems developed with technical privacy-by-design and by-default measures or technical ethics-by-design measures such as continuous bias monitoring.

RTAI procedural measures are commonly included in guidance documents. Although no single guidance document treats all of the core principles, they each treat several of them,   again showing broad agreement on the core principles from section 2. Table 2 displays prominent examples of RTAI guidance documents. Although helpful as a starting point, these guidance documents are high-level without much specificity regarding how to operationalise RTAI. This gap leaves ample opportunity to innovate by adding granular level detail to these procedures.

Guidance Documents	RTAI Values
Guidance on AI and Data Protection, The UK's Information Commissioner's Office (the data protection regulator).	Advisory document on best practices for **data protection**-compliant AI and how the ICO interprets data protection law as it applies to AI systems processing personal data. Includes the ICO's auditing process. **Transparency**, **accountability** and fairness are also data protection principles under UK GDPR.
Understanding artificial intelligence ethics and safety, The Alan Turing Institute.	Guidance intended for the public sector on development and deployment of AI and as a complement to UK Government Data Ethics Framework. Includes the principles **fairness, accountability, sustainability safety**, and **transparency**.
Catalogue of Tools & Metrics for Trustworthy AI, OECD.	Part of the OECD AI Policy Observatory, this catalogue contains tools (technical, procedural and educational) and metrics and benchmarks for **trustworthiness** across the AI lifecycle. Tools are submitted by their creators and categorised according to the OECD framework for Tools for Trustworthy AI.

Ethics-by-Design and Ethics of Use Approaches for Artificial Intelligence, The European Commission.	Guidance for research activities involving development or use of AI systems or techniques based on the experience of several AI ethics-focused EU Horizon research projects. Includes principles respect for human agency, privacy, personal data protection and data governance, fairness, individual, social, and environmental well-being, transparency, and accountability and oversight as well as steps for **ethics-by-design**.
Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment, European Commission.	A detailed assessment list (with web tool) for organisations to self-assess the **trustworthiness** of their AI systems. Derived from the principles-based Ethics guidelines of the EU's High-Level Expert Group on AI.
Artificial Intelligence Risk Management Framework, US National Institute of Standards and Technology.	A **risk management** approach intended to help organisations developing AI manage the many risks of AI systems and promote **trustworthy** and **responsible** development and use. Voluntary, sector- and use case-agnostic.
capAI Oxford Internet Institute.	A procedure for **conformity assessment** of AI systems aligned with the EU's proposed AI Act, but applicable elsewhere. Intended to support independent assessment and provide guidance on translating high-level ethics principles into verifiable criteria for **trustworthy AI**.
Portfolio of AI Assurance Techniques, DEI and Department for Science, Innovation and Technology (DSIT).	Features 14 annotated case studies of industry best practice from across a variety of sectors, exemplifying how different techniques can be used to **promote RTAI** and mapping these to the principles **safety, security & robustness, appropriate transparency & explainability, fairness, accountability and governance and contestability and redress**.
Auditing Algorithms, The Digital Regulation Cooperation Forum (DRCF).	Considers the potential role of different actors, including regulators and external parties, in the **AI assurance** ecosystem.

Table 2: Key Responsible and Trustworthy AI guidance

Table 2 shows some that of the RTAI principles are well-covered in these guidance documents, such as transparency and fairness, while others are not, such as non-maleficence. It is overly burdensome for AI developers to consult multiple guidance documents. The RTAI framework described in section 2 is intended to provide an established set of core RTAI principles for the UK. As a complement to the present report, fine-grained operational guidance would be a welcome innovation in the near future.

Other efforts attempt to operationalise the core principles by providing more practical, actionable tools – both procedural and technical – to be implemented at particular phases of the AI lifecycle. Table 3 provides a representative list of such tools organised according to the different AI lifecycle phases. In combination with the guidance documents catalogued in Table 2, these tools not only guide current AI developers to implement RTAI, but they also reveal where the current state-of-the-art ends and innovation can begin.

AI Lifecycle Phases	Examples
Assessment and planning	• AI impact assessments – a diverse field of methods for identifying positive and negative impacts to safeguard benefits and mitigate risks. Can be based on **data protection, fundamental rights, ethics,** or other perspectives, and shall be conducted throughout the AI lifecycle. • Funding programmes – The EU's Horizon Europe research funding programme makes **technical robustness** an evaluation criterion for AI projects.
Design	• **Value-sensitive design, privacy-by-design, ethics-by-design** – various methods for including principles into design processes (IEEE ethical-aligned design). • Methods for raising awareness of **data ethics, collection and use** for designers, e.g., ODI data ethics canvas. • Participatory design approaches (e.g., NESTA Participatory AI for humanitarian innovation) that bring in the **principles** of **fairness, accountability, transparency, human oversight and data protection** involving impacted stakeholders.

Development and procurement	• UK government, Office of AI, Guidelines for AI procurement – a summary of best practices focusing on **data ethics, data protection, governance and explainability, and fairness** to address specific **challenges of acquiring** Artificial Intelligence technologies in government. • World Economic Forum, AI Procurement in a Box focusing on data quality, fairness, security, human oversight and societal impact. • Data sheets for data sets – a process for documenting **datasets** used for ML tools intended for **high-risk** environments and to achieve **transparency and fairness**. • **Security** guidance (e.g., German BSI **Security of AI systems**).
Deployment, Monitoring and Control	• AI Audits, either bespoke **trustworthiness** auditing or general audits that refer to principles. E.g., **ICO AI audit**. • UK government Central Digital and Data Office and Centre for Data Ethics and Innovation's Algorithmic transparency reporting standard. • AI Incident databases – databases aiming to collect incidents of **harm** or near harm from AI systems to help researchers and developers avoid repeated unwanted outcomes.
Decommissioning/ Retirement	• There is limited stand-alone guidance for this phase, which is sometimes part of full-lifecycle guidance. Some non-AI specific guidance can be found, for example: ICO guidance personal data storage limitation and retention policies.
All Phases	• Digital Catapult Ethics Framework – a framework prompting product developers to address key questions in seven core areas of ethical AI: benefits of the service, knowing and managing risks, using data responsibly, being worthy of trust, promoting diversity, equality and inclusion, being open and understanding in communications, considering the business model. This framework is relevant to **all of the RTAI principles** from section 1.

Table 3: Examples of AI life cycle stage-specific tools, methods and guidance

As these tables show, a robust set of guidance documents and tools to implement RTAI exist. This catalogue ought to be a useful resource to AI developers seeking to operationalise some, but not all, of the core RTAI principles. In particular, data protection, robustness, transparency, fairness and security are well covered, while there are gaps in safety and non-maleficence. Even with tools where the RTAI principles are covered, there are often additional, practical steps to be considered. Consequently, as with the guidance documents, more granular operationalisation techniques are still required.

However, these methods and tools are voluntary, and their implementation can vary in quality, lacking in consistency and rigour. Furthermore, some elements of RTAI attract unbalanced attention leaving an incomplete landscape of tools and solutions. For example, there are more practical development and monitoring tools than those dedicated to design and planning. In the area of transparency and explainability, there has been a focus upon technical explainability (so that data scientists can understand how their model is working and the features that influence its outputs), rather than a **sociotechnical approach** that includes a focus on the contextual or procedural aspects (so that affected parties can understand the full context of how a decision about them was made). Within AI robustness, there is a potential over-emphasis on adversarial attacks – reflected by thousands of academic papers written on this topic in the last ten years as well as significant media attention – over the risks arising from simple failure. The OECD database tracks over 400 'tools', the plurality of which are addressed to fairness and explainability , while safety, human rights protection, and, other forms of maleficence (especially sustainability) are less represented. More granular knowledge of the nature of the different RTAI principles and their sociotechnical context could help alleviate these barriers. As AI systems are made for and by people, technological development is inextricably intertwined with human, social and environmental factors, all of which must be thoroughly understood.

Furthermore, the existence of this suite of tools can be misleading. These tools provide assistance to increase the responsibility and trustworthiness of an AI system, but they do not guarantee it, nor do they work without complementary human effort. Implementing RTAI principles requires detailed knowledge of the principles and sociotechnical understanding, in combination with data science expertise.

As a result of these gaps in RTAI guides and tools, there remains significant opportunity to innovate and develop consistent, comprehensive RTAI guides and tools across the AI lifecycle. These will be addressed after the next subsection.

BARRIERS TO RTAI

In addition to the gaps identified in the preceding subsection, the current AI landscape includes several barriers to ensuring the deployment of AI meets RTAI requirements. Dismantling these barriers requires investment and attention. Combining the gaps identified in the previous subsection with the barriers identified in this subsection provide opportunities for innovation.

These barriers exist throughout the AI lifecycle and include:

Technical Barriers	Environmental & Economic Barriers	Implementation Barriers
Outliers in datasets	Carbon & Water footprint	Lack of knowledge of sociotechnical approaches
Model degradation	Business models	Unbalanced attention given to principles
Poor Data Quality		
Substandard data collection processes		

Table 4: Barriers to RTAI

AI systems can be prone to errors owing to outlier cases being excluded from a training dataset or because conditions in the world can diverge from those in the training data. It can be difficult to predict when a machine learning model is going to err or break, and all models can be expected to lose performance over time and require retraining.

AI technologies are fundamentally reliant upon the quality of training and input data, and poor data collection and management are at the heart of many AI trust issues. Large language models in particular can be prone to 'hallucinations' – generating plausible-sounding, but factually incorrect text.

Contemporary AI also has **a high environmental impact** in terms of energy & water. For example, according to a recent study, training a large AI model such as GPT-3 can directly consume up to **700,000 litres of clean freshwater**, which is enough to produce 370 BMW cars or 320 Tesla electric vehicles. The same study also estimated that a conversation with an AI chatbot such as ChatGPT can consume up to **500 ml of water** for 20-50 questions and answers. This number must be multiplied by the 100 million active users engaging in multiple conversations. Chat GPT-4, expected to have a larger model size, is predicted to further amplify these water consumption statistics. The United Nations climate reports state that climate change is **a global emergency** reaching beyond national borders. More than a century of burning fossil fuels as well as unequal and unsustainable energy and land use have led to global warming of 1.1°C above pre -industrial levels. This has resulted in more frequent and more intense extreme weather events that have caused increasingly dangerous impacts on nature and people in every region of the world.

AI business models may also contribute to a lack of trust. Many AI systems are proprietary and provide little ability for end-users to investigate how they work. Where companies are under pressure to win business, or be first to market, the steps needed for responsible development and deployment may come under pressure. Despite high-level consensus on principles, there is **no unanimously accepted socio-ethical framework for implementing RTAI**. Different actors may interpret the core principles differently and hence develop varying solutions for implementing them into the design and management of their AI systems. AI system developers may perceive that in the current environment they can bring an unreliable product to market and still find adoption.

Organisations may lack resources for the ongoing and collaborative processes needed for operationalisation, involving cross-disciplinary team building, documentation, and knowledge management. The complexity and relative novelty of AI and an imbalance in expertise between procurers and developers of AI systems can result in the development and procurement of untrustworthy and/or irresponsible AI. Even reading, digesting and implementing each of the separate guidance documents above across the AI lifecycle represents a significant investment that can be a barrier to smaller companies, including start-ups and publicly funded organisations.

Among developers, there is still room to advance knowledge **on how to develop RTAI** or on the social and institutional requirements of trust. Furthermore, there can exist a perception that work on ethics or data protection will block scientific advances in the field if developing more powerful predictive techniques attracts more career success than devoting effort to reliability or robustness, reducing energy use, or removing bias. Reliability and robustness are not always perceived as cutting-edge topics, contributing to a potential replicability crisis in AI research.

Combining these barriers with the gaps in RTAI guides and tools shows where innovation in RATI can begin. The key RTAI innovation priorities are presented in the next subsection.

INNOVATIVE METHODS NEEDED TO ACHIEVE RTAI

While there are increasing numbers of tools and methods to support RTAI (see Tables 2 and 3), there are multiple short-term and long-term research, development and demonstration barriers to their full development (see Table 4). Requirements need to be fulfilled to address the challenges to implementing the core principles set out in section 2 and to achieve the economic and societal benefits from RTAI in the long term. A full list of innovation areas across six RTAI categories – AI assurance, ethics-by-design, trustworthiness characteristics, security, privacy & data protection and sustainability – is included in Annex I. These six categories match the areas of ongoing standardisation in RTAI, which is fully assessed in section 5. From this list of RTAI categories, **a set of priority innovation areas** has been distilled based upon **the largest potential impact in RTAI**, and the key enablers for achieving a principles-based approach in this area have been identified. Driving innovation in these areas would significantly advance the development of RTAI and position the UK as a global leader in this area.

The following innovation gaps focus on **processes and methods**, as well as the need for **understanding the context of use** of AI. This is because RTAI is a deeply **socio-technical** objective, requiring technological factors and human, social, organisational and environmental dimensions to be considered together. This is not to discount the importance of technological fundamentals, but rather to highlight how even these are **embedded in a social context**. The integration of technology and sociotechnical approaches to RTAI creates a space that is rich with innovation opportunities.

Priority innovation gaps

1 **Responsible and trustworthy data** – Quality data, adequate for the envisaged use-case is fundamental for AI. Many AI trust issues emerge from poor quality data, inappropriate proxies, biases present in historic data, or where data is collected in ways that are harmful, exploitative or invade a person's privacy. Innovations in good quality data creation and innovations in privacy-preserving methods for developing and using AI such as **federated learning, synthetic data, differential privacy** or **tools for data governance** offer a multiplier effect across RTAI.

2 **Fundamentals for AI assurance** – Innovation is needed to provide the core elements in terms of **principles-based metrics, assessment and testing methodologies, testing infrastructure, user explanations**, and **monitoring tools** needed for reliable AI assurance. Also needed is **better knowledge** of the arrangement of governance mechanisms and assurance processes that will enable AI assurance in the specific UK context.

3 **Sustainable AI** – Innovation is needed in the fundamental technological advances necessary to **reduce energy, water, and rare materials use** from AI (for example, by more **efficient computation** or **waste-heat re-use**) and **increase circularity** within hardware, as well as **benchmarks and measures** to allow comparison among AI systems of their environmental impact. Sustainability is a fundamental consideration for social trust and responsible innovation. Innovation here is necessary to support other green policies and environmental strategies and avoid undermining the environmental benefits from AI use.

4 **Sociotechnical AI professionals and research structures**. RTAI suffers from a shortage of people who can work across social and technical dimensions, and from legacy education and career models that discourage interdisciplinary and sociotechnical competencies. Innovation is needed in how we **educate and train** AI developers and users, **research and research funding structures, ethical review processes**, and **professional recognition mechanisms**. In the future, AI professionals should be able to **understand the social dimension of a technology** they are implementing and do design work in this context.

5

Exemplars and paragons of RTAI – The AI industry moves very rapidly but is still in its infancy. There are currently relatively few public, independently verified examples of RTAI done well. These examples should emerge if the other innovation gaps are filled, and the right incentives structures are in place. However, the RTAI sector needs to **capture and share innovation in RTAI** and make sure that best practices are apparent. Customers need examples of things they can request from providers, and developers need examples they can adapt or learn from. In contexts where AI developers use existing **modules, templates, or integrate other 3rd party AI systems**, they need responsible and trustworthy options to pick from. Initiatives in this space such as the OECD toolkit and CDEI portfolio of assurance techniques are very welcome and would benefit from expansion.

If supported and developed through funding or policy initiatives, these key innovation areas would significantly boost the status of RTAI in the UK, thereby advancing the position of the UK as a global leader in this area. While innovation cannot be separated from technological advancement, AI systems are built by and for humans, meaning there are human and social factors embedded in their development and use. These systems can have a significant impact on individual lives, societal wellbeing and a sustainable environment. Human and social factors in AI are often overlooked, and yet methods and tools to address them constitute the foundation of RTAI. Driving the development of these RTAI priorities would help achieve the economic and societal benefits from AI in the long term. Finally, as innovation often yields commercial opportunities, these key areas facilitate many of the commercial opportunities outlined in section 4.

BENEFIT OF THIS SECTION TO STAKEHOLDERS

For researchers and research-funding bodies, the benefit of the section is identifying areas to explore and support with research investment. For policymakers, the benefit is to understand the areas fundamental to achieving RTAI that require policy support to drive investment. Advancing these areas and methods will demonstrate the UK's nuanced understanding of AI as a sociotechnical landscape, where AI systems can have significant impact on people's lives and the environment.

Key takeaways and actionable guidance to support and develop key innovation priorities:

- 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, especially safety and non-maleficence.

- 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.

4. Commercial opportunities

Businesses in the UK can use AI to make their processes more efficient, productive, and cost-effective. 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%. At the same time, there will be major **decreases in the costs** associated with AI in the 2020s. This goes for the hardware needed to run AI systems, especially special-purpose semiconductors, the cost of energy needed for the computation of AI systems, and the costs of developing AI systems and training AI models.
The usage and implementation of AI will become easier and at the same time basic AI literacy and skills will likely expand in the work force.
These developments will lead to an acceleration in the uptake of AI systems by businesses as well as consumers, creating a positive feedback loop for the overall AI market, leading to the creation of new jobs.

By 2030, AI will be a major driver for businesses and will transform every sector of the economy.

Already at present, private-sector companies offer "AI-driven", "AI-powered" and "AI-enabled" products and services in most sectors of the economy. Alongside an element of hype,[2] real commercial opportunities abound, especially for sectors in which AI uptake has so far been relatively slow. This section 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 case studies throughout this section provide concrete examples of commercial opportunities derived from the RTAI principles.

[2] When AI is mentioned on companies' 2021 earnings calls, their share prices were forty percent more likely to increase. 'The Art of AI Maturity', Accenture.

COMMERCIAL OPPORTUNITIES FOR RTAI PRINCIPLES

This subsection catalogues the principal categories of AI systems to provide a common frame of reference for areas in which RTAI principles can be applied. Next, it shows the benefits that RTAI implementation affords and how doing so can create tangible opportunities in trust, compliance and scalability, facilitating market capture.

The following is an indicative list of commercial uses for AI, each of which includes **opportunities for RTAI adoption and implementation** to achieve full market potential.

Generative AI (GenAI) creates original text, code, images, audio, and/or video, usually based on a text prompt. Since the release of ChatGPT in November 2022, public and industry attention has focused on GenAI. It has potential applications in virtually every industry by making work more efficient and workers more productive. It is forecast to halve the time software engineers spend on coding tasks and will lower the cost of some design and creative processes to near-zero. Implementations include **AI and voice assistants** and chatbots. One market report suggests that 50% of all online content will be generated by these technologies by 2033 and predicts the global value of the market in this area to reach more than US $51bn by 2028.

GenAI must be trained on a vast amount of data requiring very high levels of energy and water. There are opportunities to integrate RTAI principles to minimise impact on privacy, intellectual property, fairness and sustainability, which will have a positive impact on market potential.

Insight generation finds patterns and anomalies in data. This can include **screening**, **forecasting**, and r**easoning with knowledge structures**. Such technology may be used to understand consumer behaviour, to identify fraud and manage financial and security risks, to analyse job or credit applicants, to assist with medical diagnoses, and to provide assurance in legal reasoning. These systems are sometimes combined with **decision support systems** that assist users with planning and executing responses.

Datasets can contain billions of data points. Algorithmic models can unintentionally draw on inputs that correlate with protected characteristics, such as ethnicity and gender, creating biased and unfair output and resulting in poor market uptake. Identifying bias and regularly training algorithms can support organisations to capitalise on market opportunities.

Recommender systems are used in advertising, retail, streaming services and social media. The global market for Content Recommendation Engines, which had a valuation of $2bn in 2020, is expected to increase to $33bn by 2028, growing at more than 40% per year. AI-based **personalisation** allows organisations to automatically tailor outputs to individual users, such as advertising, educational content, and even medical treatments. **Network optimisation** systems are used to plan distribution networks for logistics services or public utilities.

To make personalised recommendations, these systems must input personal data from the targeted consumer potentially violating the right to **privacy** and transgressing **data protection** laws. Addressing this could create positive brand recognition and build consumer trust, thereby enabling organisations to capitalise on the economic benefit of these tools.

Robots, drones, and CAVs (connected and automated vehicles) apply AI systems in physical environments, including in manufacturing, warehouse automation, and precision agriculture, including surveillance, crop inspection, equipment-monitoring, and even transport. According to a 2021 analysis, the global market for self-driving cars was estimated to be worth $22bn and is projected to grow to $76bn by 2027, expanding annually at a rate of more than 22%. AI systems applied to **multimodal control systems** enable the automation of traffic control systems, building services (such as HVAC and agricultural systems).

Automated decision-making in complex environments can blur lines of **accountability**, thereby hindering a person's right of **redress** when **harmed**, and in turn, lowering market value. Increased transparency and human-in-the-loop provisions can identify harms and provide a mechanism for addressing them proactively.

Specific **recognition** systems are required to make sense of disorganised inputs like images and language. **Machine vision** systems can label images and video, recognise faces, analyse medical images, and satellite imagery. **Natural language processing** (NLP) systems identify patterns in texts and speech and are used in machine translation, legal discovery tools, automatic transcription, sentiment analysis and market research.

In these systems, **bias** can arise due to the quality of data used, **privacy** concerns arise when these systems are used to analyse personal data such as emails or social media posts. Building in sufficient **transparency and explainability** can build trust by providing information on how these systems work and drawing consumers towards these products.

By implementing the core RTAI principles in each of the above AI implementation areas, businesses can capture the following benefits which will facilitate market capture for UK businesses within and beyond UK borders.

Trust. Organisations want to drive AI at scale to reap the benefits of automation and data driven analysis. Implementing RTAI principles (whether internally or though assurance products and services described in the next subsection) is necessary to create trust that the insights that AI systems produce are not biased or built on illicitly collected personal data. This will build an organisation's customer base by driving trust within the marketing in its products and services.

Compliance. The deployment of RTAI will involve compliance with existing regulations and standards. But, more importantly, it places organisations at the forefront of a developing policy landscape enabling companies to meet their *pre-emptive* needs for compliance and influencing the direction of market rules and norms. Furthermore, implementing the RTAI principles, which section 2 showed to be globally acceptable, will make UK companies' offerings primed for major markets in Asia, North and South America.

Scalability. Companies can show strategic reasons to generate trust in their work and in their products. Accenture identifies a positive correlation between RTAI and revenue growth in AI companies, picking out a small group of 12% of organisations who generate 50% more revenue growth, and who are 53% more likely to apply responsible AI practices. The increased revenue may be a result of customer loyalty, brand reputation, employee engagement, and ethical business practices, drawing upon and contributing to CSR practices that are globally scalable.

Combining these benefits with the above-mentioned areas where there is clear economic opportunity for commercial use of RTAI will compound the potential gains of UK companies looking to move into this market. This will create a positive feedback loop that builds these companies' reputation and further encourages uptake of their products and services. It will also offset the risks related to a lack of trust and responsibility.

CASE STUDY: QUANTUM COMPUTING

An independent report on the Future of Compute (2023), as commissioned by the Department for Culture, Media and Sport (DCMS) and announced in the National AI Strategy: AI Action Plan (2022), reviewed the future strategic direction for the range of technologies under the umbrella term of 'compute'. 'Compute' refers 'to computer systems where processing power, memory, data storage and network are assembled at scale to tackle computational tasks beyond the capabilities of everyday computers.' The review found that the UK's existing **compute capabilities** are unfit to meet the needs of AI users, and risk falling behind those of other advanced economies. Whilst the report itself has been criticised for missing an opportunity to establish clear governance and ensure responsible use of compute, particularly in the context of high-risk AI development, the government has responded swiftly to the recommendation that there should be an immediate and significant increase in large-scale accelerator-driven compute for AI research by announcing a £900ml investment into the creation of a new 'exascale' computer and a dedicated AI Research Resource. Advancing this market has an RTAI opportunity for the government to outline a vision for ensuring that the UK's capacity-building and ultimate use of compute will be managed sustainably. AI computations are energy-intensive, and the quality of AI results tends to improve with increased computation. As a result, AI systems have a large carbon footprint and by 2030 the information and communication sector could account for up to 51% of global electricity consumption.

One market opportunity to raise compute sustainably can be found in quantum computing, which can potentially help business meet their **ESG goals** without compromising performance. Quantum computing is still in its infancy, but can theoretically solve certain problems with far less energy, including many computationally-expensive problems. While development is still needed to reduce the energy consumption of quantum computing, it has the potential to power AI systems that are both high-performing and more **environmentally sustainable**.

CASE STUDY: SYNTHETIC DATA

Artificially generated information, known as synthetic data, replicates the characteristics of real-world data but can be used without revealing sensitive information. A synthetic data set has the same mathematical properties as the real-world data set it is replacing, but it does not contain the same information. It is generated by taking a relational database, creating a generative machine learning model for it, and generating a second set of data.

The result is a data set that contains the general patterns and properties of the original.

Synthetic data can be used to test machine learning models or build and test software applications without compromising real, personal data. In a global market where data is one of the most valuable resources, an infinite amount of data can potentially be produced quickly, cheaply, and safely. The Synthetic Data Generation Market was valued globally at $168.9ml in 2021, and one analysis predicts 35.8% annual growth by 2031, when it is expected to reach $3.5bn. According to a widely referenced Gartner study, 60% of all data used in the development of AI will be synthetic rather than real by 2024.

Driving a reliable and safe synthetic data market would directly support the implementation of RTAI principles of **privacy and fairness**.

AI ASSURANCE MARKET

Organisations that want to enhance trust or demonstrate compliance with RTAI principles or legislation will need access to a network of commercial offerings focused on RTAI assessments, audits and certifications. This subsection demonstrates the significant opportunities to be leveraged in the still latent AI assurance ecosystem. It concludes with an assessment of why a fragmented AI assurance ecosystem currently exists in the UK and strategies to resolve prevailing challenges.

While companies widely advertise AI capabilities, products, and services, uptake of 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 2022 report 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.[3]

As such, there is a clear commercial opportunity for AI assurance products and services. In a 2022 study, the AI governance market was forecast to be worth more than $1bnglobally by 2026, growing annually at a rate of 65%. However, AI governance is just a part of the RTAI market. It is reported that at least 80% of companies will commit more than 10% of their AI budget to meet regulatory requirements by 2024, and 45% expect to spend at least 20% of their AI budget on regulatory issues. If half of that is spent on pure compliance activities that do not encompass responsibility and trust, one might expect that just over 6% of AI budgets are spent on activity related to RTAI more widely. If in 2024 the AI market is worth £24bn,[4] then the RTAI market (including internal investment within companies) is worth £1.5bn.

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

[4] Drawing on https://www.gov.uk/government/publications/ai-activity-in-uk-businesses/ai-activity-in-uk-businesses-executive-summary. See also DCMS/Capital Economics, 'AI Activity in UK Business' .

CASE STUDY: RTAI IN RECRUITMENT

The UK recruitment sector, which was worth £43bn in 2022, has been quick to see the value of AI. There is now a scramble to adopt LLMs in generating job advertisements and analysing applicant materials. However, there is a low recognition of the benefits of AI assurance in recruitment. There is strong potential for applying fairness tools in recruitment, especially as they relate to bias, to avoid mistakes such as Amazon's quickly-abandoned 2018 recruitment tool that discriminated against women. Given such cautionary tales – as well as new legislation such as the European AI Act and the New York City bias audit law – heavy demand can be expected for tools that can assure workforces, applicants, regulators, and customers that recruitment tools will avoid bias. Several firms in the UK now offer services for bias auditing. Given that technical bias research is relatively well-developed, including in open source, further opportunities can be expected to provide bias assurance in recruitment tools and related applications by properly packaging existing techniques. Furthermore, experts warn that technical debiasing is not the whole solution. There are different metrics to measure bias that can differ significantly and are appropriate to different contexts. The variety of metrics exacerbates risks of "ethics-washing" if unscrupulous organisations report the most flattering metrics rather than the most appropriate ones. Accordingly, alongside these technical tools, there is a large space in this sector for sociotechnical, qualitative impact assessment work.

Leveraging these market opportunities would advance the RTAI principles of **fairness, accountability** and **appropriate transparency**.

Consulting firms provide services for examining AI systems and analysing whether they are compliant, responsible and trustworthy. As the risks of AI systems increase with their complexity, scalability and adoption by businesses and society, there is tremendous potential for RTAI and for AI assurance services. By assuring the AI systems of UK companies and in exporting assurance services and techniques abroad, responsible service providers can provide evidence of good practice that will support their growth and scalability within and beyond the UK. The list below provides a description of AI assurance areas that will have the greatest impact on UK businesses and consumers.

AI Assurance Technical Services	AI Assurance Consultancy Services
Fairness metrics	Human Rights, Democracy, Rule of Law (Huderia) Impact Assessments
Explainability metrics	Fundamental Rights Impact Assessments
Algorithmic re-training	Algorithmic Impact Assessments
Transparency metrics	Conformity Assessments
Data cleaning	Bias Audits
Performance stability	Compliance Audits
Accuracy metrics	RTAI Policy & Procedure Development
Cybersecurity	RTAI training

Table 5: List of AI Assurance services

As the table above demonstrates there are two types of assurance services that need to be developed. **AI assurance technical solutions** include technical means applied to AI systems to ensure responsibility and trustworthiness. For example, there are numerous analytics products that focus on measuring explainability or bias. **AI assurance consultancy services** include services to assess the impact of algorithms and advise developers on RTAI principles such as fairness, human rights and compliance. It is essential to advance both in order to provide a holistic assurance ecosystem that analyses and improves RTAI across the whole AI lifecycle.

To mature these services and capture the potential market opportunities, the **UK assurance market for RTAI will have to address the following challenges**:

Lack or ignorance of standards/ regulation. Standards are lacking or still in development for many AI applications. However, even where there are standards or regulation, many UK businesses and organisations are uncertain which standards and regulations apply. For example, the CDEI found that developers of Connected and Autonomous Vehicles (CAVs) were widely concerned about a lack of standards, despite hundreds of available standards catalogued in centralised and publicly-available databases. Standards organisations will have to agree and publicise core standards to support the assurance industry and businesses and assurance providers will need to educate themselves on the standards' content and scope.

Ignorance of regulation and standards can be addressed by creating and disseminating knowledge resources (such as the AI Standards Hub), and other documents about assurance (such as this report and the CDEI Roadmap documents). (See section 5 for a full assessment of UK policies and standards relevant to RTAI.)

Uneven demand from assurance users. Across sectors and principles, the demand for assurance services is varied, despite common public anxieties about AI. The UK assurance markets for privacy, data protection, safety, security and robustness (including data quality) are relatively mature, with the cybersecurity sector as a whole worth £10.5bn. By contrast, assurance regarding other principles is newer and is sought less consistently. For example, independent assurance is rarely sought for AI applications in human resources (HR), such as software for hiring or promoting employees, despite the high risks regarding bias and fairness, privacy and data protection, and accountability.

UK government can address the uneven demand for assurance through the introduction or clarification of new regulation and standards. A mixture of cross-sectoral and sectoral regulations can protect public well-being and stimulate greater trust in AI.

A complementary way to resolve the uneven demand for assurance is to provide support, especially for start-ups and SMEs, in the form of accelerators, sandboxes, regional hubs and academic partnerships. Such support can provide training and resources for developing RTAI where they are most needed.

Lack of common approaches to assurance. Assurance providers take different approaches to assessing responsibility and trustworthiness. This is not a problem if appropriate techniques are used for each application, for example if security compliance is assured through certification and non-maleficence is assessed through impact assessment. However, different assurance providers might assess accountability and governance by examining differing kinds of evidence and employing disparate assessment metrics. The result of this sort of diversity can be uncertainty on the part of developers about how to develop RTAI, and uncertainty on the part of assurance providers about how AI systems should be assessed or compared.

AI assurance providers ought to develop and publicise good practice with which standards bodies can align. The UK Government could consider **the establishment of a dedicated AI authority** that can support these strategies, perhaps in partnership with the ongoing efforts of the CDEI.

The standards and regulations that can support the assurance industry are described in section 5.

GLOBAL MARKET LEADERSHIP IN ASSURANCE

Harnessing these **commercial opportunities**, the UK is poised to become a market leader in AI assurance globally by exporting knowledge, products, services and practices.

Regarding **knowledge**, UK higher education institutions (HEIs) and businesses already produce research, training and thought leadership that have global reach. Enhancing research and training in RTAI through the key innovation priorities in section 3 will strengthen policy and industry initiatives, and directly serve to export practices in the form of training and research. In industry, assurance practices can cross borders through products, services and the policies of international companies. UK **policies** can influence RTAI assurance practices globally (and RTAI innovation more generally) by leading by example in creating demonstrably effective guidelines, strategies, policies and contributing to international standards (see section 5). Moreover, access to the UK market will require suppliers to conform to UK requirements, creating an international amplification effect for UK RTAI. The UK can also help spread practices around the world by supporting international civil society work and policy initiatives, such as the Global Partnership on AI, and multistakeholder initiatives that encourage interaction between knowledge, industry and policy actors (Figure 4).

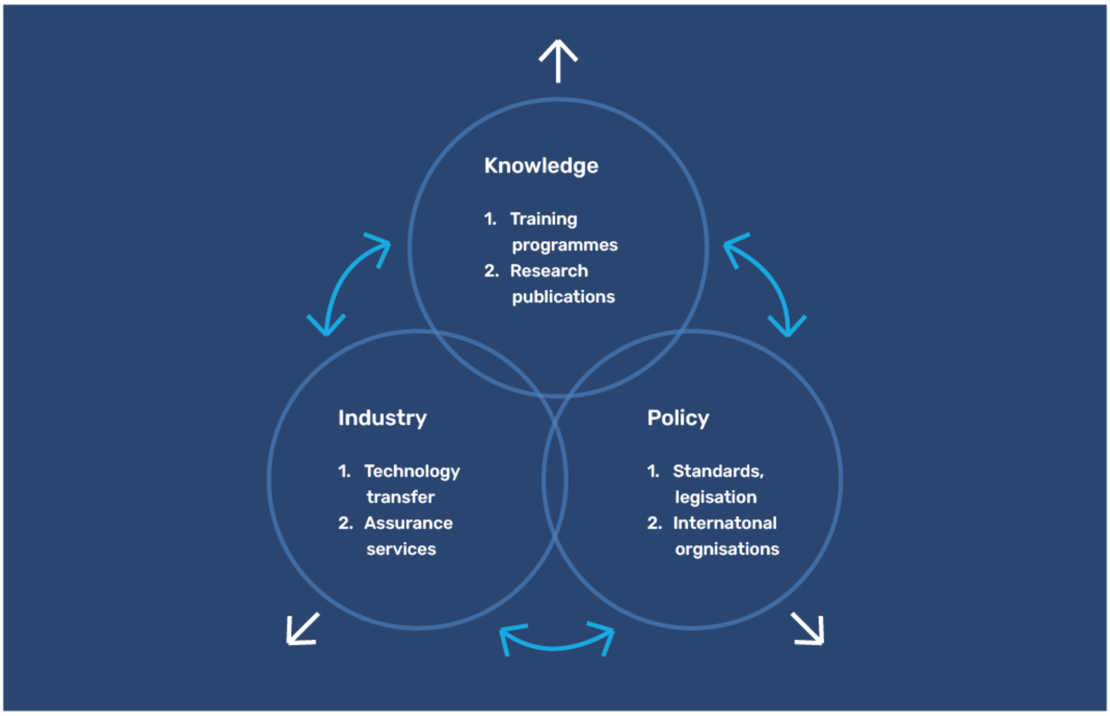

Figure 4: Methods for the UK to export assurance practices

Finally, the UK can provide a leading example on the global stage by:

- Continuing to invest in research, including funding for universities, research institutions, and private companies working on AI assurance.
- Swiftly and effectively implementing regulatory and support mechanisms across industry, thereby leading by example for other jurisdictions.
- Making RTAI alignment a requirement for all public sector procurement of AI tools and services.
- Contributing to standards and supporting the involvement of stakeholders who lack resources to contribute independently, such as start-ups and small NGOs and enterprises.

- Including RTAI in other standards and policies at early stages.
- Supporting multistakeholder initiatives that bring together public, private and third sector organisations to develop practical RTAI resources and processes.
- Coordinating and supporting international certification schemes.

BENEFIT OF THIS REPORT TO STAKEHOLDERS

This guide to industry stakeholders identifies new avenues of development and growth in the UK deriving from RTAI in order to benefit from economic and social advantages of RTAI in the long term. It presents actionable insights on the uses of AI and commercial opportunities for RTAI and the assurance industry in order to facilitate market expansion and market capture.
This section provides references to other information sources so that stakeholders can do further research on particular uses of AI, industry sectors, risks and assurance methods.

Key actionable guidance to expand and leverage commercial opportunities in AI:

- 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.

5. Policy and standards analysis

Recent policy and standard development activities in the UK demonstrate a trajectory towards increased domestic support for research and development investments and expansion of international leadership on the global RTAI stage. In September 2021 , following the central recommendation of the AI Council in the AI Roadmap, the UK government published the National AI Strategy, which outlines a 10-year vision for maximising the commercialisation and exploitation of AI, benefitting from high economic and productivity growth due to AI, and most ambitiously, **establishing the most trusted and pro-innovation system for AI governance in the world**.
UK government published a White Paper entitled A pro-innovation approach to AI regulation ("AI White Paper"), to further explicate the UK's proposed approach. In its response to the AI White Paper, the Information Commissioner's Office (ICO) has expressed support for the creation of a **central supervisory function** to deliver on a range of commitments, including the proposed creation of a **multi-regulator sandbox** for AI as recommended in Sir Patrick Vallance's Pro-Innovation Regulation of Technologies Review (2023). Looking into the near future of AI development, the Competition and Markets Authority (CMA) has also announced an Initial Review (2023) into AI foundation models, including LLMs and generative AI.

Supportive regulation through **policies** and **standards** can cement the UK's position as a global leader in RTAI. This section outlines policy approaches from other countries and international organisations to foster alignment in international cooperation and trade. In doing so, it identifies where the UK can influence international activities in RTAI. It also gives an overview of existing standards and ongoing standardisation efforts in the UK and internationally to show where the UK can continue to advance its domestic RTAI objectives. The main findings from this section demonstrate that the UK can drive domestic and international 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 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.

INTERNATIONAL POLICIES ON RESPONSIBLE AND TRUSTWORTHY AI

Identifying opportunities for alignment between UK policies and international initiatives will facilitate the export of UK products and services and drive UK global leadership in RTAI. This subsection describes current international policies directly relevant to UK RTAI and identifies how the UK can leverage its strong commitment to RTAI to influence international guidelines and polices.

United States (US): The UK and US have "A Shared Vision for Driving Technological Breakthroughs in Artificial Intelligence", including recognising the importance of promoting trust and understanding in AI.

The shared vision between the UK and US aims to:

- Take stock of and utilise existing bilateral science and technology cooperation (e.g., the Memorandum of Understanding between the U.S. National Science Foundation and UK Research and Innovation on Research Cooperation) and multilateral cooperation frameworks;
- Recommend priorities for future cooperation, particularly in R&D areas where each partner shares strong common interest (e.g., interdisciplinary research and intelligent systems) and brings complementary challenges, regulatory or cultural considerations, or expertise to the partnerships;
- Coordinate, as appropriate, the planning and programming of relevant activities in these areas, including promoting researcher and student collaboration that could potentially involve national partners, the private sector, academia, and the scientific community to further our efforts by harnessing the value of public-private partnerships; and
- Promote research and development in AI, focusing on challenging technical issues, and protecting against efforts to adopt and apply these technologies in the service of authoritarianism and repression.

By implementing and actioning the core RTAI principles, key innovation priorities and commercial opportunities identified in this report, the UK can lead the bilateral achievement of these stated objectives. Particularly, the key innovation priorities set out in section 3 can be used to recommend priorities for future cooperation, facilitate private-public partnerships and protect against efforts to apply AI systems in the service of authoritarianism and oppression. The operationalisation of the core RTAI principles will likewise help achieve this latter goal. By adopting the guidance set out in this report, the UK has opportunities to influence the US (and other international partners) and continue to demonstrate its commitment to RTAI.

Additional US frameworks encouraging trustworthy AI include the 2023 updated National AI Research and Development Strategic Plan, "The Blueprint for an AI Bill of Rights", Executive Order 13960 (public AI use cases registry), and a federal call on the US National Institute of Standards and Technology (NIST) by Congress for developing an AI assurance roadmap, including development metrics, assessment tools and technical standards . These other instruments further align with UK RTAI principles seeking to promote trust and transparency in AI systems, and in supporting an AI assurance ecosystem.

In addition to the US framework and potential partnership there are other areas where the UK can influence international cooperation on RTAI. The Organization for Economic Cooperation and Development, UNESCO and the European Union have RTAI initiatives and principles with which the UK can explore synergies and drive cooperation.

OECD: Recommendation on Artificial Intelligence (first set of intergovernmental principles for trustworthy AI) and OECD AI Policy Observatory.

As a member of OECD, the UK can lead the international formulation and implementation of RTAI principles. As stated in section 2, the core RTAI principles in this report are fully aligned with those supported and propagated by the OECD. UK developments and exemplars of good practice in meeting RTAI requirements, including meeting the innovation and investment priorities and commercial opportunities identified in sections 3 and 4, can be disseminated through the OECD to have global impact among its members. As a result, the UK can position itself to lead on international cooperation for RTAI.

UNESCO: UNESCO's Recommendation on the Ethics of Artificial Intelligence adopts a human rights-based approach and asks countries to pay special attention to the needs of Low-to-Middle Income Countries (LMICs), including Least Developed Countries (LDCs), Landlocked Developing Countries (LLDCs) and Small Island Developing States (SIDS). The RTAI principles in section 2 are aligned with UNESCO's objectives for ethical AI. As a member of the U.N., the UK can leverage the RTAI framework and innovation opportunities in this report to lead the global community to achieve UNESCO's aims to protect human dignity and well-being and prevent harm.

EU: The EU has three interrelated **legal initiatives intended to contribute to RTAI**: a legal framework for AI to address fundamental rights and safety risks of AI systems (the so-called forthcoming AI Act), a civil liability framework to update existing liability rules for AI, and revisions to various sectoral safety legislation.

Upon its entry into force, the AI Act will also **apply to UK providers** that place AI systems on the market, or put them into service, in the EU, as well as UK providers and users of AI systems if the output produced by the system is used in the EU. As such accessing the EU market, which is one of the UK's largest markets, will require alignment between UK and EU requirements in these key areas:

- Alignment on how responsibility and liability for demonstrating compliance with AI regulatory principles will or should be allocated to existing supply chain actors (such as chip developers, data and computational resources providers, and end-users) within the AI lifecycle.

- Alignment on assessing and mitigating the environmental footprint of AI systems and foundation models. With respect to the sustainability, the UK has an opportunity to influence the development and uptake of environmental requirements and standards. Leading on matters of AI **sustainabilit**y would fulfil the core principle of non-maleficence and demonstrate global leadership in an urgent area of concern.

- Alignment on the functioning of the **AI assurance market**, including the content and role of technical and procedural standards, risk assessments and enforcement bodies in providing assurance.

In fact, influence on standards is a core way in which the UK can direct alignment through positive cooperation rather than regulatory enforcement. The following section identifies opportunities to capture and challenges to tackle in relation to standards in order for the UK to lead in RTAI implementation.

UK STANDARDISATION EFFORTS IN RESPONSIBLE AND TRUSTWORTHY AI

UK standardisation efforts concerning RTAI are increasing. These efforts are essential to provide much-needed guidance to organisations creating AI systems as well as to organisations providing AI assurance services. While the UK utilises its relationship with the US and other OECD countries to set international standards for RTAI implementation, it also has an opportunity to leverage these relationships to influence European standards, given the **legal agreements between UK and EU** standardisation bodies.

The **UK occupies a strong leadership position** in the international standards system as a founder of the standards organisations the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). A significant number of important international standards committees are managed by the UK. Through standards, UK businesses, including SMEs, and British consumers shape decisions that are taken by businesses on every continent.

With respect to legally required cooperation with Europe, EU AI Act envisages **a substantial role for standards bodies** in drafting technical standards to support key technical areas covered by the Act. The European Commission has already issued a draft standardisation request to the European standardisation organisations in support of requirements for design and development of high-risk AI systems, AI provider's **quality management systems, conformity assessment and auditing** of AI systems and r**obustness specifications**. This activity has **implications for the UK**. As participants in CEN and CENELEC, BSI representatives contribute to developing European standards which means **UK experts will be able to influence implementation of guidelines for international trade to the EU**. One area where the UK has already led is the algorithmic transparency reporting standard (still in a consultation phase), published by the Central Digital and Data Office (CDDO) and Centre for Data Ethics and Innovation (CDEI) in 2021. Its purpose is to support transparency regarding the use of AI in the public sector.

One of the functions for technical standards identified in the white paper "A Pro-Innovation Approach to AI Regulation" is to provide guidelines on **methods to assess, design, and improve transparency, explainability, and fairness**. CEN/CLC JTC 21 have taken the position that standards on harm, risk and trustworthiness will need stakeholders from a variety of backgrounds involved at the development stage. Their concern is shared by some critics of AI standards, including in the Ada Lovelace Institute's Inclusive AI Governance discussion paper which recommends more inclusive representation within the national standards body. The UK has a significant opportunity to **demonstrate its commitment to the RTAI principles by fostering increased representation**. It can achieve this objective through funding and training for stakeholders representing interests outside of traditional areas of industry and academia. As lack of diversity and ensuring representation are issues across the globe, a robust UK initiative in this area can be highly influential.

One of the foreseen applications of technical standards in the EU, and across the globe, is to underpin **impact assessments, auditing and performance testing for AI assurance**. This scope of application creates a significant sphere of influence for UK standards. The CDEI suggests that various techniques underlined by technical standards, including impact and risk assessments, and algorithmic audits, are important for supporting the assurance of AI products relative to a range of AI risks and use cases. The AI Standards Hub has built a database of standards currently in development and already published from various standards development organisations to help practitioners navigate the fast-developing AI standards landscape.

Nevertheless, current standardisation efforts would benefit from including **measurement standards** needed for compliance and assurance practices (CDEI, 2021). For example, **establishing thresholds for bias audits or impact assessments** is essential to avoid a proliferation of empty assurance services that leave consumers confused or even lead to ethics-washing. RTAI standards are widely recognised as needed and yet they remain under-developed internationally. This situation provides favourable circumstances for the UK to provide genuine leadership through its sincere commitment to developing actionable RTAI standards.

SCORING UK POLICY AND STANDARDS

In order to provide guidance to UK policymakers and standards bodies, this section scores the development maturity of UK RTAI policies and standards in relation to the principles identified in section 2, the key innovation priorities in section 3, and the commercial opportunities identified in section 4, according to the following metrics (Table 6):

Score	Description
0	**Non-existent:** No policy or standard in evidence.
1	**In development:** Policy or standard in draft form and/or some informed practice.
2	**Existent:** Policy or standard finalised, evidence of application.
3	**Operationalised:** Policy fully implemented across different sectors. Note: This score does not concern standards because publication of a standard is its last step, thus, there is no equivalent measure for operationalising since standards are voluntary.

Table 6: Scoring metrics for assessing the maturity levels of the core principles, innovation gaps and commercial opportunities related to RTAI

Principles (see section 2)	Score	Qualitative comments
Appropriate transparency and explainability	**1/2**	**Policy**: Proposed in the 2022 Policy Paper and included in the AI White Paper (2023), together with a definition and rationale. As it relates to data protection law, the principle of transparency in AI is fully operationalised in the UK GDPR, compliance with which is overseen by the ICO. Broader application requires further implementation by other regulators within their sectors and remits.
	0/1	**Standards:** No technical standards exist that address explainability, but some standards for transparency in the context of trustworthiness are in development: e.g., ISO/IEC WD 12792, and the CDDO/CDEI national algorithmic transparency standard is at a consultation phase.
Safety, security and robustness	**2**	**Policy**: Underpinned by product safety laws, proposed in the 2022 Policy Paper and included in the AI White Paper (2023), together with a definition and rationale, this principle is to be fully operationalised by regulators within their remits.
	1/2	**Standards**: Guidelines are in development: CEN/CLC NWIP Trustworthiness characteristics as mandated by the EC standardisation request in support of the EU AI Act will address, among others robustness and security specifications. Plenty of sector specific safety guidelines for robots, drones and CAVs exist. ISO/IEC, IEEE, ETSI, BSI and ITU-T have published standards for information security governance and cybersecurity framework development. ISO/IEC AWI 27090 will seek to address cybersecurity threats within AI specifically.

Non-maleficence	**1**	**Policy:** Implicit in, and underpinned by, existing non-AI-specific laws and policies with which the design, development and use of AI systems is required to comply. In addition, the AI White Paper (2023) includes an initial assessment of AI-specific risks and their potential to cause harm to human rights and other protected values. However, it is also stated that a range of societal challenges are outside the scope of the proposed regulatory framework, including the issue of environmental sustainability.
	1	**Standards:** Technical standards for risk management have been published by IEEE, ISO/IEC and NIST, with impact assessment standards still in development. CEN/CLC is expected to develop their own risk assessment guidelines to support the AI Act, but these are at pre-draft stage.
Fairness and justice	**1/2**	**Policy:** Proposed in the initial policy paper (2022) and included in the AI White Paper (2023), together with a definition and rationale, fairness is fully operationalised in data protection law as a principle of the UK GDPR. Broader application, however, depends on further implementation by regulators within their sectors and remits. As a corollary to fairness, justice is in principle effected by the requirement for the design, development and use of AI systems to comply with existing laws.
	1	**Standards:** Technical standards to support treatment of unwanted bias have been published by the BSI, IEEE and ISO/IEC. Harmonised standards addressing ethical and societal concerns supporting the implementation of the EU AI Act will be published by CEN/CLC but are currently at pre-draft stage.

Privacy and data protection	**2**	**Policy:** This principle is embedded into the regulatory framework via data protection law, wherein the UK regulator, the ICO, has taken an active role in its operationalisation, for example by issuing Guidance on AI and data protection and publishing an AI and data protection risk toolkit. Also subsumed in principles of fairness, safety, security and robustness in the AI White Paper (2023), and thus to be further implemented by other regulators within their sectors and remits.
	2	**Standards:** Standards that address privacy and data protection in AI have been published by BSI, ETSI, IEEE and ISO/IEC.
Accountability	**1/2**	**Policy:** The version proposed in the 2022 Policy Paper has been further refined and included in the AI White Paper (2023), together with a definition and rationale. A cross-cutting principle of data protection law under the UK GDPR, the more wide-ranging application of this principle across the AI lifecycle requires further implementation by other regulators within their sectors and remits. Notably, it is recognised in the AI White Paper (2023) that there is a lack of clarity around the appropriate allocation of responsibility and liability to different supply chain actors within the AI lifecycle, with the government planning to consult with a range of experts to further its understanding on this issue.
	1	**Standards:** Most standardisation towards institutionalising trust building practices through governance and system management pertains to IT in general. Some AI specific guidelines are in development, e.g., ISO/IEC CD 42001, currently in draft form.

Innovation opportunities (see section 3)	Score	Qualitative comments
AI assurance related	2	**Policy:** The AI White Paper (2023) emphasises the critical role of assurance techniques in enabling responsible adoption of AI. The government plans to launch a Portfolio of AI Assurance Techniques in June 2023 to complement existing resources, such as the CDEI's Roadmap to an effective AI assurance ecosystem (2021) and AI Assurance Guide.
	0/1	**Standards:** Process standards such as system management are in development while performance standards cannot be developed before measurements are agreed upon. Some performance benchmarking standards have been published by IEEE.
Sociotechnical AI	1	**Policy:** As part of the wider collection of guidance on using AI in the public sector, the Alan Turing Institute report Understanding AI Ethics and Safety provides a guide for the responsible design and implementation of AI systems in the public sector. It complements the Data Ethics Framework, a practical guide for appropriate and responsible use of data in government and the public sector.
	1	**Standards:** ISO/IEC TR 24368 provides an overview of ethical and societal concerns in relation to AI. No methods for addressing these concerns horizontally have been developed, though some sector specific standards have been published by BSI and IEEE, e.g., for automated vehicles and robots.

Responsible and trustworthy data	**2**	**Policy:** In its response to the Data: A new direction consultation, the government highlights the need for continued investment into research and development for privacy-enhancing technologies (PETs), as well as further guidance tailored to the public in order to build trust in the use of PETs and complement existing resources, such as the ICO draft guidance on anonymisation and PETs, and the CDEI's PET Adoption Guide. The UK is also collaborating with the US on prize challenges to advance the use of PETs to combat financial crime. Priorities for Pillar One of the National AI Strategy include improving access to good quality, representative data. As highlighted in the AI Action Plan (2022), government actions against this aim include publishing a Policy Framework for enabling responsible data sharing in line with Mission One of the National Data Strategy, as well as supporting the Open Data Institute's (ODI) Data Assurance work programme, the aim of which is to ensure the trustworthiness of data and data practices.
	2	**Standards:** Many data protection guidelines exist, e.g., ETSI GR SAI 002 V 1.1.1, and for specific sectoral applications, e.g., for PAS 186:2020 smart cities. There are proposals to develop guidelines specific to big data security and privacy as it pertains to AI in IEEE. CEN/CLC has been mandated by the EC with developing standards for specifications on transparency, robustness, accuracy and data quality. ISO/IEC has published many standards on data quality, including the 8000 series. Some sector specific standards for application have been published by IEEE and ANSI.
Sustainable AI	**0**	**Policy:** Although advocated by some stakeholders in response to the 2022 Policy Paper, sustainability is not one of the five principles listed in the AI White Paper (2023). The Future of Compute Review emphasises the need for government action to build sustainable compute capabilities in order to limit the environmental impacts of this enabling technology.
	0	**Standards:** Proposals for standards that quantify the environmental impact of AI have been started at both ISO and CEN, however no horizontal guidance is in development as of yet.

Commercial opportunities (see section 4)	Score	Qualitative comments
Generative AI	**2**	**Policy:** The review on pro-innovation regulation for digital technologies (2023) recommends that the government adopts a clear policy position on the relationship between intellectual property law and generative AI in order to provide confidence to innovators and investors. In its Response, the government highlights the ongoing work of the Intellectual Property Office to achieve this, including plans to publish by the summer (2023) a code of practice on the use of copyrighted works to support the interests of both the AI and creative industries. Furthermore, as announced in the Integrated Review, there are plans to establish a government-industry Foundation Model Taskforce, to which an initial £100 million in funding has been pledged.
	0	**Standards:** No standards that specifically support the implementation of responsible and trustworthy generative AI exist.
Insight generation (e.g., decision support systems)	**1**	**Policy:** There are some examples of AI systems being used by NHS health and care organisations to assist with medical diagnoses, including by analysing brain scans and x-ray images.
	0	**Standards:** CSA has published CAN/CIOSC 101 for ethical design and use of automated decision systems, and IEEE has proposed developing guidelines for procurement of decision support systems.

Recommender systems (e.g., personalisation and network optimisation)	**1**	**Policy:** AI and Machine learning (ML) network optimisation tools have been trialled in the energy sector by both the UK Power Network and the National Grid. In 2022, the ICO launched a stakeholder consultation on the use of AI/ML recommendation systems to protect people from content-related harm.
	0	**Standards:** No standards exist to support implementing recommender systems.
Robots, drones and connected and automated vehicles (CAVs)	**2**	**Policy:** The Government Response to the review on pro-innovation regulation for digital technologies (2023) highlights its role in advancing the series of recommendations relating to drones, including by supporting the Future Flight Challenge, the projects associated with which are centred around public good use cases and include drone-based distribution of medicines in Scotland (CAELUS). The government's vision for supporting the safe deployment of self-driving vehicles through a new legislative framework is laid out in the Connected and Automated Mobility 2025 policy paper (2022). It is supported by the CDEI's report on Responsible Innovation in Self-Driving Vehicles, which builds on the joint report by the Law Commissions and examines how these legislative proposals can be supported ethically through trustworthy and responsible regulation, governance, and assurance.
	2	**Standards:** Terminology, safety, testing and performance evaluation standards have been published by ISO/IEC, IEEE, ITU-T and the BSI.

Recognition systems (e.g., machine vision, NLP)	**1/2**	**Policy:** Based on the use cases for machine vision outlined in section 4 relevant UK policy includes the research, development and use of AI imaging in health and social care, as well as the growing use of automated facial recognition (AFR) technology by police. The College of Policing has issued guidance on the use of live facial recognition by police, while the British Security Industry Association has issued more general guidance on the ethical and legal use of AFR. An example of the similarly emerging use of NLP in the public sector is the deployment of such techniques by the Government Digital Service to improve accessibility of information on GOV.UK.
	1/2	**Standards:** Horizontal standards for NLP and machine vision performance are still at early development stages. ISO/IEC has published on biometrics performance and IEEE has published imaging standards for the healthcare sector.

BENEFIT OF THIS REPORT TO STAKEHOLDERS

To advance its position as a global leader in RTAI, the UK can continue to develop policies and standards that support the core principles and advance the innovation priorities and commercial opportunities described in this report. To foster international cooperation and trade, and to show where the UK can advance its international influence, this section identifies collaborative opportunities with the US, OECD and the EU. Subsequently, this section gives an overview of existing standards and ongoing standardisation efforts so that UK policymakers can clearly identify which RTAI-relevant areas need further development. This section concludes that AI assurance and sustainable AI are two areas providing opportunities for the UK to show its sincere commitment to RTAI and to influence international policies and standards in these areas.

Key actionable guidance:

- 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.

6. Conclusion

This report serves as a single and common frame of reference on core principles, key innovation priorities and new commercial opportunities relating to responsible and trustworthy AI (RTAI) in the UK. 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 this critical and fast-paced area. 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.

Annex I: Innovation opportunities in responsible and trustworthy AI

Based upon the identification of key innovation priorities in section 3, this annex contains the full list of areas where there are opportunities for innovation to support principles-based approaches to responsible and trustworthy AI. In the following technological, organisational and social areas, there is opportunity and need for innovation that would benefit from active funding and research.

AI assurance – tools and methods to allow for reliable impact assessment, conformity assessment, governance, and management of AI systems, underpinning any AI assurance ecosystem (a set of tools and services that allow users and other interested parties to know that AI are effective, trustworthy and legal). Including:

- Methods for holistic evaluation of AI systems and their complex social contexts, including documenting intended use-cases, red-teaming, pre-training/pre-deployment risk assessments, consultation with affected parties, and accountable design.
- Training, professional recognition, and career pathways for qualified assessors of the implementation and evaluation of AI ethics and responsible AI development.
- Transparent, widely adopted frameworks for 3rd party conformity assessment.
- Social science research into the arrangement of professional societies, peer/self-governance mechanisms, codes of conduct and institutions that would be best suited to help determine appropriateness of practices in RTAI development, assess difficult cases, identify negligent behaviour, and sanction bad actors.
- Research on effective AI governance structures, including into how codes of responsible professional behaviour in AI can best be embedded and actively enforced in organisational leadership and cultures.

- AI governance and monitoring tools for ML models in production at scale. Tools that support organisations to manage compliance requirements across large numbers of deployed AI systems.

Ethics-by-design – the concrete integration of ethics into the design of AI systems.

- Examples of RTAI, validated by independent third parties. These would serve as shareable examples of best practice, that can serve as models for deployment or the basis for further refinement. Particularly valuable for the public sector.

- Empirically proven methods to translate principles into design strategies or patterns in real-world development contexts, across a range of different sectors. These will need updating as AI techniques advance, e.g., a privacy-preserving fraud prevention model in finance.

- Methods for empirical verification of ethical design, e.g., methodologies for field-trials and pilots of AI systems.

- Adequate funding and support for interdisciplinary research that appropriately combines technical and engineering skills with social science, ethics, and domain knowledge expertise. This support can be more expensive than single discipline research and risks falling between discipline-based funding models.

- Interdisciplinary ethics review processes that go beyond harms to research participants to encompass social impacts and are accessible and implementable outside academia as well as within. Ways to support diverse stakeholders to contribute their perspective and trust requirements into AI development.

Trustworthiness innovation – tools and techniques that improve features of AI systems that reinforce trustworthiness and offer improvements to how RTAI is developed in practice.

- Benchmarks for the different dimensions of RTAI principles (transparency, fairness, privacy protection, what being non-biased means in a particular context of use, etc.) and methods for formal validation of elements of trustworthiness. Guidance on selecting appropriate benchmarks from amongst these.

- Testing infrastructure, sandboxes and test data sets, including those that test on edge- and unusual cases and the various particularly challenging areas of principles.

- Improved approaches to transparency, interpretability and explainability. Despite claims around the “black box” nature of deep learning, black boxes are a single design choice, and explainability for AI developers themselves can be acceptable. However, the requirements of explainability vary for different audiences such as trained end-users, untrained end-users, regulators and other people affected by automated decisions, meaning that approaches to explainability need to be developed that meet all these various needs.

- Foundational computer science research that can potentially lead to alternative paradigms in AI research – For example, new and refined approaches to machine learning that prioritise robustness, reliability, safe failure modes and back-up plans, transparency, and accountability – potentially over raw predictive power. Like a field of science, AI has trends and approaches that gain traction on particular problems. These trends, and the design and problem-solving methods associated with them can change over time. The field should remain open to new ways of thinking about AI problems, to avoid path dependency.

- Software engineering methodologies supporting responsible innovation, e.g., integration of responsibility principles into machine learning operations (MLOps). MLOPS is a widely adopted approach to the reliable and efficient deployment and maintenance of machine learning models in production, including practices for collaboration and communication between data scientists and operations professionals.

- Data quality is a necessary precondition of RTAI, therefore efforts to increase data quality (not necessarily data volume), representativeness and accuracy in well-managed, legally and ethically collected data sets will contribute to trustworthy AI be ensuring that systems are trained and used on good quality data, as will tools for identifying anomalies in data.

- Methods to calculate, quantify and communicate uncertainty and confidence levels.

- Training and communication methods to increase AI literacy in individual sectors where trustworthy and responsible AI is required to unlock innovation and uptake (see section 4).

- Training and development of professional practitioners of RTAI system development and implementation.

- RTAI templates for non-specialist developers adding AI elements to other applications. Methods to independently verify reliability of AI subsystems and components.

Security – Security innovations can improve the reliability of AI systems and prevent other principles being maliciously undermined.

- Integration of information security best practices into machine learning development (for example secure storage and encryption of data used for training or testing, access controls to development environments, separation of privileges or incident management).

- Identification of the main adversarial vulnerabilities of AI models (including generative AI) that cause unreliability and innovative methods to avoid or mitigate these, including methods to distinguish between malicious attacks and errors.

- Methods for assessing the risk and potential impacts of adversarial attacks on different concrete AI applications.

- Improvements in ML techniques that contribute to security, such as regularisation, generalisation, adversarial retraining, decreased model outputs, data sanitisation, and avoiding overfitting. These can help prevent AI systems being unduly impacted by errors or deliberate poisoning of data used for training or evaluation.

- Trusted and verified training sources for model retraining and transfer learning. Transfer learning is an approach to machine learning that repurposes a model training on one task for a second related task, with the intent of reducing the resources needed for training. If the initial model includes exploitable vulnerabilities these may be inherited by the retrained model.

Privacy and data protection – Innovation with the potential to reduce potential harms to privacy from the data collection and processing necessary for training and operating AI.

- Robust and repeatable methods and technologies for safe, fair, legal and ethical collection, processing, storage and sharing of data, and datasets used to train AI systems, including ensuring the provenance of data and tracking consent for use of personal data in AI training.

- Technological approaches to enhance privacy protection such as anonymisation, differential privacy, homomorphic encryption, federated machine learning and secure multi-party computation, or synthetic data, as well as improvements in ease-of-implementation and efficiency of these tools to remove current barriers to uptake.
- Good quality guidance on implementing such technologies, including mapping privacy-enhancing technologies to appropriate use cases and legal concerns.

Sustainability – Potential innovations to reduce the environmental impact of AI.

- Methods and metrics for assessing and comparing the environmental benefits and costs of AI systems (e.g., energy use, water use, carbon output, rare earth materials, ecosystem impacts).
- Innovation in more environmentally sustainable and resource-aware AI fundamentals (energy efficient hardware and software, reducing computation requirements, re-use of heat waste), methods of design for sustainability, and processes to encourage their uptake, e.g., processes and ecosystems for hardware recycling and circularity.

Annex II: Current standards relevant to responsible and trustworthy AI

Publisher	Standard code	Standard name	Development stage	Principle/ opportunity addressed	Horizontal / sector specific application
CDDO/CDEI	N/A	Algorithmic transparency standard	Consultation	Transparency	Public sector
ISO/IEC	WD 12792	Information technology — Artificial intelligence — Transparency taxonomy of AI systems	Draft	Transparency	Horizontal
ISO/IEC	WD TS 5471	Information technology — Artificial intelligence (AI) — Quality evaluation guidelines for AI systems	Draft	Transparency	Horizontal
CEN/CLC	NWIP	AI trustworthiness characterization	Pre-draft	Transparency	Horizontal
IEEE	7001	IEEE Standard for Transparency of Autonomous Systems	Published	Transparency	Horizontal
ISO	16300-3	Natural language description for abstract scenarios for automated driving systems. Specification	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
IEEE	P2976	Standard for XAI – eXplainable Artificial Intelligence – for Achieving Clarity and Interoperability of AI Systems Design	Pre-draft	Explainability	Horizontal
IEEE	P2894	Guide for an Architectural Framework for Explainable Artificial Intelligence	Pre-draft	Explainability	Horizontal
ISO/IEC	AWI TS 6254	Objectives and approaches for explainability of ML models and AI systems	Pre-draft	Explainability	Horizontal
ISO/IEC	DIS 25059	Information technology — Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems	Draft	Safety	Horizontal

ISO	TR 4804	Road vehicles. Safety and cybersecurity for automated driving systems. Design, verification and validation	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
ITU-T	Y.4471	Functional architecture of network-based driving assistance for autonomous vehicles	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
CSA	CAN/CIOSC 101	Ethical design and use of automated decision systems	Published	Ethics-by-Design	Insight generation
VDI	VDI-EE 4030	Consideration of human reliability in the design of autonomous systems	Published	Safety	Horizontal
ANSI	ANSI/RIA R 15.08-1	Industrial Mobile Robots – Safety Requirements – Part 1: Requirements for the Industrial Mobile Robot	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
ANSI	ANSI/UL 4600	Standard for Safety for the Evaluation of Autonomous Products	Published	Safety	Horizontal
ISO	TR 21934-1	Road vehicles. Prospective safety performance assessment of pre-crash technology by virtual simulation – State-of-the-art and general method overview	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
ISO	22733-1	Road vehicles. Test method to evaluate the performance of autonomous emergency braking systems. Car-to-car	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
BSI, ISO	BS ISO 39003	Road Traffic Safety (RTS). Guidance on safety ethical considerations for autonomous vehicles	Published	Ethics-by-Design	Robots, drones and connected and automated vehicles (CAVs)
CEN	CEN/TS 17395:2019	Intelligent transport systems. eSafety. eCall for automated and autonomous vehicles	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
IEEE	2846	Assumptions in Safety-Related Models for Automated Driving Systems	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
BSI, CEN	BS EN ISO 13482:2014	Robots and robotic devices. Safety requirements for personal care robots	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)

BSI	BS 8611:2016	Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems	Published	Ethics-by-Design	Robots, drones and connected and automated vehicles (CAVs)
IEEE	P7009	Fail-Safe Design of Autonomous and Semi-Autonomous Systems	Published	Accountability	Robots, drones and connected and automated vehicles (CAVs)
IEEE	P7008	Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems	Published	Ethics-by-Design	Robots, drones and connected and automated vehicles (CAVs)
ITU-T	F.749.13	Framework and requirements for civilian unmanned aerial vehicle flight control using artificial intelligence	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
ISO/IEC	TR 24028	Overview of trustworthiness in artificial intelligence	Published	Trustworthiness	Horizontal
ISO/IEC	AWI TR 5469	Artificial intelligence — Functional safety and AI systems	Pre-draft	Safety	Horizontal
ETSI	GR SAI 005 V 1.1.1	Securing Artificial Intelligence (SAI) – Mitigation Strategy Report	Published	Security	Horizontal
ETSI	GR SAI 002 V 1.1.1	Securing Artificial Intelligence (SAI) – Data Supply Chain Security	Published	Data protection	Horizontal
ETSI	GR SAI 001 V 1.1.1	Securing Artificial Intelligence (SAI) – AI Threat Ontology	Published	Security	Horizontal
ISO/IEC	15026-2:2011	Systems and software engineering. Systems and software assurance. Assurance case	Published	Robustness	Horizontal
ETSI	GR SAI 006 V 1.1.1	Securing Artificial Intelligence (SAI) – The role of hardware in security of AI	Published	Security	Horizontal
BSI	PAS 186:2020	Smart cities. Supplying data products and services for smart communities. Code of practice	Published	Data protection	Smart cities
IEEE	P2945	Standard for Technical Requirements for Face Recognition Systems	Pre-draft	Security	Recognition systems

ISO/IEC	TR 24741	Information technology – Biometrics – Overview and application	Published	Privacy	Recognition systems
IEEE	2842	IEEE Recommended Practice for Secure Multi-Party Computation	Published	Cybersecurity	Horizontal
ISO/IEC	30145-3	Smart City ICT reference framework. Smart city engineering framework	Published	Privacy	Smart cities
BSI, ISO-IEC	BS ISO/IEC 24661	Information technology. User interfaces. Full duplex speech interaction	Published	Security	Generative AI
BSI, ISO	BS ISO 37166	Smart community infrastructures. Urban data integration framework for smart city planning (SCP)	Published	Data protection	Smart cities
ISO/IEC	TS 27570	Privacy protection. Privacy guidelines for smart cities	Published	Data protection	Smart cities
ISO/IEC	TS 27110	Information technology, cybersecurity and privacy protection. Cybersecurity framework development guidelines	Published	Cybersecurity	Horizontal
ISO/IEC	27022:2021	Information technology. Guidance on information security management system processes	Published	AI assurance	Horizontal
IEEE	P3129	Standard for Robustness Testing and Evaluation of Artificial Intelligence (AI)-based Image Recognition Service	Draft	Robustness	Recognition systems
ITU-T	M.3080	Framework of artificial intelligence enhanced telecom operation and management (AITOM)	Published	AI assurance	Generative AI
ITU-T	F.748.13	Technical framework for the shared machine learning system	Published	Security	Horizontal
ISO/IEC	20547-3	Information technology – Big data reference architecture – Part 3: Reference architecture	Published	Security	Horizontal
IEEE	2830	Technical Framework and Requirements of Trusted Execution Environment based Shared Machine Learning	Published	Data protection	Horizontal
ETSI	GR SAI 004 V 1.1.1	Securing Artificial Intelligence (SAI) – Problem Statement	Published	Fairness	Horizontal

BSI	BS PAS 1885:2018	The fundamental principles of automotive cyber security. Specification	Published	Cybersecurity	Robots, drones and connected and automated vehicles (CAVs)
BSI	BS PAS 1881:2022	Assuring the operational safety of automated vehicles. Specification	Published	Safety	Robots, drones and connected and automated vehicles (CAVs)
BSI	BS PAS 1880:2020	Guidelines for developing and assessing control systems for automated vehicles	Published	AI assurance	Robots, drones and connected and automated vehicles (CAVs)
ISO/IEC	27014:2020	Information security, cybersecurity and privacy protection. Governance of information security	Published	Cybersecurity	Horizontal
ISO/IEC	25023	Systems and software engineering. Systems and software Quality Requirements and Evaluation (SQuaRE). Measurement of system and software product quality	Published	Trustworthiness	Horizontal
ISO/IEC	23751:2022	Information technology. Cloud computing and distributed platforms. Data sharing agreement (DSA) framework	Published	Data protection	Horizontal
ISO/IEC	20547-4:2020	Information technology. Big data reference architecture. Security and privacy	Published	Privacy	Horizontal
ISO	22166-1	Robotics. Modularity for service robots. General requirements	Published	Security	Robots, drones and connected and automated vehicles (CAVs)
ANSI/CTA	2090	The Use of Artificial Intelligence in Health Care: Trustworthiness	Published	Trustworthiness	Healthcare
NIST	AI RMF	NIST Artificial Intelligence Risk Management Framework	Published	Non-maleficence	Horizontal
ISO/IEC	WD 9868	Remote biometric identification systems — Design, development, and audit	Pre-draft	Robustness	Recognition systems

IEEE	P3157	Recommended Practice for Vulnerability Test for Machine Learning Models for Computer Vision Applications	Pre-draft	Robustness	Recognition systems
ISO/IEC	TR 29119-11	Software and systems engineering – Software testing – Part 11: Guidelines on the testing of AI-based systems	Published	Robustness	Horizontal
ISO/IEC	TR 24029-1	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview	Published	Robustness	Horizontal
ISO/IEC	23894	Information technology – Artificial intelligence – Risk management	Published	Non-maleficence	Horizontal
ISO/IEC	DIS 24029-2	Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods	Draft	Robustness	Horizontal
ISO/IEC	AWI 42005	Information technology — Artificial intelligence — AI system impact assessment	Pre-draft	Non-maleficence	Horizontal
IEEE	7010	IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being	Published	Non-maleficence	Horizontal
CEN/CLC	N148	AI-enhanced Nudging	Pre-draft	Non-maleficence	Horizontal
ISO/IEC	NP TS 17847	Information technology – Artificial intelligence – Verification and validation analysis of AI systems	Pre-draft	AI assurance	Horizontal
ISO/IEC	TR 24714-1	Information technology – Biometrics – Jurisdictional and societal considerations for commercial applications – Part 1: General guidance	Published	Non-maleficence	Recognition systems
ISO	AWI 13482	Robotics — Safety requirements for service robots	Pre-draft	Safety	Robots, drones and connected and automated vehicles (CAVs)
IEEE	7000	IEEE Standard Model Process for Addressing Ethical Concerns during System Design	Published	Ethics-by-design	Horizontal
IEEE	2813	Big Data Business Security Risk Assessment	Published	Non-maleficence	Horizontal
IEEE	P2817	Guide for Verification of Autonomous Systems	Pre-draft	Trustworthiness	Horizontal

ISO/IEC	AWI TS 29119-11	Information technology — Artificial intelligence — Testing for AI systems — Part 11	Published	Non-maleficence	Horizontal
BSI	BS PAS 11281:2018	Connected automotive ecosystems. Impact of security on safety. Code of practice	Published	Non-maleficence	Robots, drones and connected and automated vehicles (CAVs)
ISO	9241-220	Ergonomics of human-system interaction. Processes for enabling, executing and assessing human-centred design within organizations	Published	Ethics-by-Design	Horizontal
IEEE	P7003	Algorithmic Bias Considerations	Published	Fairness	Horizontal
ISO/IEC	WD 19795-10	Information technology — Biometric performance testing and reporting — Part 10: Quantifying biometric system performance variation across demographic groups	Draft	Fairness	Recognition systems
ISO/IEC	TR 22116	Information technology – A study of the differential impact of demographic factors in biometric recognition system performance	Published	Fairness	Recognition systems
BSI	Flex 236 v1.0:2022-01	Enabling the development of inclusive standards Understanding the role of data and data analysis. Guide	Published	Fairness	Horizontal
IEEE	P2863	Recommended Practice for Organizational Governance of Artificial Intelligence	Pre-draft	Accountability	Horizontal
ISO/IEC	TR 24027	Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making	Published	Fairness	Horizontal
ISO/IEC	AWI TS 12791	Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks	Pre-draft	Fairness	Horizontal
ANSI	ANSI/CTA 2096	Guidelines for Developing Trustworthy Artificial Intelligence Systems	Published	Trustworthiness	Horizontal
CEN/CLC	N256	Green and sustainable AI	Pre-draft	Sustainability	Horizontal
CEN/CLC	N147	Overview of AI tasks and functionalities related to natural language processing	Pre-draft	Trustworthiness	Recognition systems

IEEE	2937	IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems	Published	AI assurance	Horizontal
IEEE	2671	IEEE Standard for General Requirements of Online Detection Based on Machine Vision in Intelligent Manufacturing	Published	Privacy	Recognition systems
IEEE	7007	IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems	Published	Ethics-by-Design	Robots, drones and connected and automated vehicles (CAVs)
ISO/IEC	AWI 27090	Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems	Pre-draft	Cybersecurity	Horizontal
ISO/IEC	TR 20226	Information technology — Artificial intelligence — Environmental sustainability aspects of AI systems	Pre-draft	Sustainability	Horizontal
ISO/IEC	19944-2	Cloud computing and distributed platforms – Data flow, data categories and data use – Part 2: Guidance on application and extensibility	Published	Data protection	Horizontal
IEEE	7002	IEEE Standard for Data Privacy Process	Published	Privacy	Horizontal
BSI, ISO-IEC	BS ISO/IEC 29155-4:2016	Systems and software engineering. Information technology project performance benchmarking framework. Guidance for data collection and maintenance	Published	Privacy	Horizontal
IEEE	P2975	Standard for Industrial Artificial Intelligence (AI) Data Attributes	Pre-draft	Data protection	Horizontal
ISO/IEC	19944-1	Cloud computing and distributed platforms. Data flow, data categories and data use. Fundamentals	Published	Data protection	Horizontal
BSI	BS 10102-2:2020	Big data. Guidance on data-intensive projects	Published	Data protection	Horizontal
BSI	BS 10102-1:2020	Big data. Guidance on data-driven organizations	Published	Accountability	Horizontal
IEEE	7005	IEEE Standard for Transparent Employer Data Governance	Published	Accountability	Horizontal

ISO	37170	Smart community infrastructures. Data framework for infrastructure governance based on digital technology in smart cities	Published	Data protection	Smart cities
ISO/IEC	TS 38505-3	Information technology. Governance of data. Guidelines for data classification	Published	Accountability	Horizontal
IEEE	3652.1	IEEE Guide for Architectural Framework and Application of Federated Machine Learning	Published	Accountability	Horizontal
ISO/IEC	22624	Information technology. Cloud computing. Taxonomy based data handling for cloud services	Published	Data protection	Horizontal
ISO/IEC	TR 38505-2	Information technology – Governance of IT – Governance of data – Part 2: Implications of ISO/IEC 38505-1 for data management	Published	Accountability	Horizontal
ISO/IEC	TR 38502	Information technology – Governance of IT – Framework and model	Published	Accountability	Horizontal
BSI, ISO/IEC	BS ISO/IEC 38507:2022	Information technology. Governance of IT. Governance implications of the use of artificial intelligence by organizations	Published	Accountability	Horizontal
BSI	BS 13500:2013	Code of practice for delivering effective governance of organizations	Published	Accountability	Horizontal
ISO/IEC	DIS 42001	Information technology — Artificial intelligence — Management system	Draft	AI assurance	Horizontal
ISO/IEC	FDIS 24668	Information technology — Artificial intelligence (AI) — Process management framework for big data analytics	Draft	AI assurance	Horizontal
IEEE	P2937	Standard for Performance Benchmarking for AI Server Systems	Draft	AI assurance	Horizontal
CEN/CLC	NWIP	AI Risk catalogue	Pre-draft	AI assurance	Horizontal
ISO/IEC	24368	Information technology — Artificial intelligence — Overview of ethical and societal concerns	Published	Ethics-by-Design	Horizontal

IEEE	P2247.4	Recommended Practice for Ethically Aligned Design of Artificial Intelligence (AI) in Adaptive Instructional Systems	Pre-draft	Ethics-by-Design	Horizontal
IEEE	P7014	Standard For Ethical Considerations in Emulated Empathy in Autonomous and Intelligent Systems	Pre-draft	Ethics-by-Design	Horizontal
ISO/IEC	DTS 4213	Information technology — Artificial intelligence (AI) — Assessment of machine learning classification performance	Draft	Trustworthiness	Horizontal
IEEE	P7011	Standard for the Process of Identifying and Rating the Trustworthiness of News Sources	Draft	Trustworthiness	Recognition systems
ITU-T	Y.3602	Big data – Functional requirements for data provenance	Published	Data quality	Horizontal
IEEE	2801	IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence	Published	Data quality	Healthcare
ISO/IEC	38505-1	Information technology. Governance of IT. Governance of data – Application of ISO/IEC 38500 to the governance of data	Published	Data quality	Horizontal
ISO	8000 series	Data quality	Published	Data quality	Horizontal
ISO/IEC	WD 29794-5	Information technology — Biometric sample quality — Part 5: Face image data	Draft	Data quality	Recognition systems
ISO/IEC	WD TS 24358	Face-aware capture subsystem specifications	Pre-draft	Data quality	Recognition systems
IEEE	P3123	Standard for Artificial Intelligence and Machine Learning (AI/ML) Terminology and Data Formats	Draft	Data quality	Horizontal
ISO/IEC	DIS 8183	Information technology — Artificial intelligence — Data life cycle framework	Draft	Data quality	Horizontal
ISO/IEC	AWI 5259-5	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance	Pre-draft	Data quality	Horizontal

ISO/IEC	AWI 5259-3	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines	Pre-draft	Data quality	Horizontal
ISO/IEC	AWI 5259-2	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures	Pre-draft	Data quality	Horizontal
ISO/IEC	AWI 5259-1	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples	Pre-draft	Data quality	Horizontal
ISO/IEC	25030	Systems and software engineering. Systems and software quality requirements and evaluation (SQuaRE). Quality requirements framework	Published	Data quality	Horizontal
ISO/IEC	25020	Systems and software engineering. Systems and software Quality Requirements and Evaluation (SQuaRE). Quality measurement framework	Published	Data quality	Horizontal
ISO/IEC	25024	Systems and software engineering. Systems and software Quality Requirements and Evaluation (SQuaRE). Measurement of data quality	Published	AI assurance	Horizontal
ISO/IEC	25012:2008	Software engineering. Software product quality requirements and evaluation (SQuaRE). Data quality model	Published	Data quality	Horizontal
ITU-T	L.1040	Effects of information and communication technology-enabled autonomy on vehicles longevity and waste creation	Published	Sustainability	Robots, drones and connected and automated vehicles (CAVs)
ITU-T	Y.4470	Reference architecture of artificial intelligence service exposure for smart sustainable cities	Published	Sustainability	Smart cities

Table 7: List of standards for RTAI

References

Ahmed, M. (8 October 2020), UK passport photo checker shows bias against dark-skinned women, BBC News, https://www.bbc.co.uk/news/technology-54349538

Arandia, P., Ley, M., Geiger, G., Méndez, M., Braun, J., Márquez, R., Constantaras, E., Howden, D., Jorrín, J., Fernández, R., & Villarino, A., (17 April 2023) Spain's AI Doctor, Lighthouse Reports, https://www.lighthousereports.com/investigation/spains-ai-doctor/

Autoriteit Persoonsgegevens (17 July 2020), Werkwijze Belastingdienst in strijd met de wet en discriminerend, https://autoriteitpersoonsgegevens.nl/actueel/werkwijze-belastingdienst-in-strijd-met-de-wet-en-discriminerend

Beauchamp, T. L., & Childress, J. F. (2012). Principles of Biomedical Ethics (7th edition ed.). Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada. https://doi.org/10.1145/3442188.3445922

Berditchevskaia, A., Malliaraki, E., & Peach, K. (September 2021) Participatory AI for Humanitarian innovation: A Briefing paper, https://media.nesta.org.uk/documents/Nesta_Participatory_AI_for_humanitarian_innovation_Final.pdf

Bietti, E. (2021). From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics. Journal of Social Computing, 2(3), 266-283. https://doi.org/10.23919/JSC.2021.0031

Bradford, A. (2019) The Brussels Effect: How the European Union Rules the World. Oxford University Press.

Brookes, J. (8 September 2021) Robotdebt was technology 'beta testing' on most vulnerable citizens, InnovationAus.com, https://www.innovationaus.com/robodebt-was-technology-beta-testing-on-most-vulnerable-citizens/

Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. UCDL Rev., 51, 399.

Central Digital & Data Office, (5 January 2023) Algorithmic Transparency Recording Standard – Guidance for Public Sector Bodies, UK Government, https://www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard/algorithmic-transparency-recording-standard-guidance-for-public-sector-bodies

CNIL (20 October 2022) Facial recognition: 20 million euros penalty against Clearview AI, https://www.cnil.fr/en/facial-recognition-20-million-euros-penalty-against-clearview-ai

Dainow, B. & Brey, P. (25 November 2021) Ethics by Design and Ethics of Use Approaches for Artificial Intelligence, Version 1.0, European Commission, https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

Declan & Gordijn, Bert. (2018). Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations. Science and Engineering Ethics. 24. 10.1007/s11948-017-9961-8.

Department for Science Innovation & Technology, & Office for Artificial Intelligence. (2023). A pro-innovation approach to AI regulation. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M. L., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. arXiv preprint arXiv:2305.02231.

Druce, J., Niehaus, J., Moody, V., Jensen, D., & Littman, M. L. (2021). Brittle AI, Causal Confusion, and Bad Mental Models: Challenges and Successes in the XAI Program. arXiv. https://doi.org/10.48550/arXiv.2106.05506

Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. Information & Communications Technology Law, 31(1), 123-153. https://doi.org/10.1080/13600834.2021.1958860

European Commission, https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal

Federal Office for Information Security (2022), Security of AI-Systems: Fundamentals, Adversarial Deep Learning, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Security-of-AI-systems_fundamentals.pdf?__blob=publicationFile&v=4

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3518482

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, A., Mokander, J., & Yuni, W. (2022) capAI – a Procedure for Conducting Conformity Assessment of AI systems in Line with the EU Artificial Intelligence Act, https://dx.doi.org/10.2139/ssrn.4064091

Garante per la protezione dei dati personali (9 March 2022) Facial recognition: Italian SA fines Clearview AI Eur 20 Million Bans use of biometric data and monitoring of Italian data subjects, https://www.gpdp.it/home/docweb/-/docweb-display/docweb/9751323#english

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé, H., and Crawford, K. (2021). Datasheets for datasets. Communications of ACM 64:12, 86–92. https://doi.org/10.1145/3458723

Geiger, et al, (6 March 2023), Supicion Machines, Lighthouse Reports, https://www.lighthousereports.com/investigation/suspicion-machines/

Gu, X., Tianqing, Z., Li, J., Zhang, T., Ren, W., & Choo, K.-K. R. (2022). Privacy, accuracy, and model fairness trade-offs in federated learning. Computers & Security, 122, 102907. https://doi.org/https://doi.org/10.1016/j.cose.2022.102907

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hao, K., & Hernández, A. P. (2022). How the AI industry profits from catastrophe. MIT Technology Review.

Heaven, D. (2019, October 9). Why deep-learning AIs are so easy to fool. Nature. https://www.nature.com/articles/d41586-019-03013-5

Heaven, W.D. (12 November 2020), AI is wrestling with a replication crisis, MIT Technology Review, https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/

Hellenic Data Protection Authority (13 July 2022) Decision 35/2022, https://www.dpa.gr/sites/default/files/2022-08/35_2022%20anonym_EN_FINAL.pdf

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. Science and Engineering Ethics, 21(3), 619-630. https://doi.org/10.1007/s11948-014-9565-5

Hickok, M. (2022). Public procurement of artificial intelligence systems: new risks and future proofing. AI & Society. https://doi.org/10.1007/s00146-022-01572-2

High Level Expert Group on Artificial Intelligence (8 April 2019), Ethics Guidelines for Trustworthy AI, European Commission, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M. & Wallach, H. (2018) Improving fairness in machine learning systems: What do industry practitioners need? ArXiv181205239 Cs. doi:10.1145/3290605.3300830

House of Representatives of the States General, (17 December 2020), Unprecedented Injustice, https://www.houseofrepresentatives.nl/sites/default/files/atoms/files/verslag_pok_definitief-en-gb.docx.pdf

Information Commissioner's Office (15 March 2023), Guidance on AI and data protection, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/

Information Commissioner's Office (18 May 2022), Enforcement Notice, https://ico.org.uk/media/action-weve-taken/enforcement-notices/4020437/clearview-ai-inc-en-20220518.pdf

Information Commissioner's Office (Undated), A Guide to ICO Audit: Artificial Intelligence (AI) Audits, https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf

Information Commissioner's Office (undated), Principle (e): Storage limitation, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/the-principles/storage-limitation/

International Association of Privacy Professionals & FTI Consulting (January 2023) Privacy and AI Governance Report: Privacy, Quo Vadis – will you lead the way? https://iapp.org/media/pdf/resource_center/privacy_ai_governance_report.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Joint Research Council (2020) Robustness and explainability of artificial intelligence https://publications.jrc.ec.europa.eu/repository/handle/JRC119336

Kaye, K. (11 July 2022) Not my job: AI researchers building surveillance tech and deepfakes resist ethical concerns. Protocol. https://www.protocol.com/enterprise/ai-computer-vision-cvpr-ethics

Kim, Eun-Sung. (2020). Deep Learning and Principal-agent Problems of Algorithmic Governance: The New Materialism Perspective. Technology in Society. 63. 101378.10.1016/j.techsoc.2020.101378.

Leslie, D., (2019) Understanding artificial intelligence ethics and safety: a guide for responsible design and implementation of AI systems in the public sector, Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529

Li, P., Yang, J., Islam, M.A., Ren, S. (2023) Making AI Less "Thirsty": Uncovering and Address the Secret Water Footprint of AI

Models. https://doi.org/10.48550/arXiv.2304.03271

Lohn, A. J. (2020). Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance. arXiv. https://doi.org/10.48550/arXiv.2009.00802

Mitchell, B. (4 January 2021) UK recover needs Computer Science graduates who are competent and ethical, Higher Education Policy Institute, https://www.hepi.ac.uk/2021/01/04/uk-recovery-needs-computer-science-graduates-who-are-competent-and-ethical/

Mittelstadt B. (2019) Principles alone cannot guarantee ethical AI. Nature Machine Intelligence 1 (11):501–507. doi: 10.1038/s42256-019-0114-4.

Morgan, R. A., van Zoonen, W., & ter Hoeven, C. (2023) Lost in the crowd? An investigation into where microwork is conducted and classifying worker types. European Journal of Industrial Relations, 0(0), 09596801231171997. https://doi.org/10.1177/09596801231171997

Murgia, M. (2019, July 23, 2019). AI's new workforce: the data-labelling industry spreads globally. Financial Times.

National Highway Traffic Safety Administration (11 April 2023), Part573 Safety Recall Report, https://static.nhtsa.gov/odi/rcl/2023/RCLRPT-23V085-9893.PDF

National Institute of Science and Technology, Facial Recognition Vendor Test (FVRT), https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

National Institute of Standards and Technology (January 2023), Artificial Intelligence Risk Management Framework (AI RMF 1.0) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

OECD (2019). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. https://oecd.ai/en/ai-principles

OECD (28 June 2021) Tools for Trustworthy AI: a framework to compare implementation tools for trustworthy AI systems. https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm

OECD, Catalogue of Tools & Metrics for Trustworthy AI, https://oecd.ai/en/catalogue/overview

Office of Artificial Intelligence (2020), Guidelines for AI procurement: A summary of best practices addressing specific challenges of acquiring Artificial Intelligence in the public sector, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/

attachment_data/file/990469/Guidelines_for_AI_procurement.pdf

Open Data Institute (2021), Data Ethics Canvas, https://www.theodi.org/article/the-data-ethics-canvas-2021/#1674123368990-c995b7bf-3325

O'Shaughnessy, M., Sheehan, M. (2023). Lessons From the World's Two Experiments in AI Governance. Carnegie Endowment for International Peace. https://carnegieendowment.org/2023/02/14/lessons-from-world-s-two-experiments-in-ai-governance-pub-89035

Partnership on AI, AI Incidents Database, https://partnershiponai.org/workstream/ai-incidents-database/

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. AI and Ethics. https://doi.org/10.1007/s43681-023-00258-9

Raji, D., Kumar I.E, Horowitz, A., and Selbst, A., (2022), The Fallacy of AI Functionality. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 959–972. https://doi.org/10.1145/3531146.3533158

Rao, R. (9 May 2022), The Dutch Tax Authority was felled by AI – What comes next?, IEEE Spectrum, https://spectrum.ieee.org/artificial-intelligence-in-government

Rességuier, Anaïs, and Rowena Rodrigues. AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics. Big Data & Society7, no. 2 (1 July 2020): 2053951720942541. https://doi.org/10.1177/2053951720942541

Reijers, Wessel & Wright, David & Brey, Philip & Weber, Karsten & Rodrigues, Rowena & O'Sullivan,

Roberts, H., Babuta, A., Morley, J., Thomas, C., Taddeo, M., & Floridi, L. (2022). Artificial Intelligence Regulation in the United Kingdom: A Path to Global Leadership? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4209504

Rudin, C. & Radin, J. (22 November 2019), Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition, Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Stahl, B., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z., & Wright, D. (2023). A systematic review of artificial intelligence impact assessments. Artificial Intelligence Review. 1-33. 10.1007/s10462-023-10420-8.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019), Ethically Aligned Design: A vision for prioritising human well-being with autonomous and intelligent systems (Version 2). https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. Nature Machine Intelligence, 2(1), 10-12. https://doi.org/10.1038/s42256-019-0136-y

UNESCO (2021) Recommendation on the Ethics of Artificial Intelligence. SHS/BIO/REC-AIETHICS/202. https://unesdoc.unesco.org/ark:/48223/pf0000380455

UNESCO, OECD, IDB (2022). The Effects of AI on the Working Lives of Women. https://unesdoc.unesco.org/ark:/48223/pf0000380861

Williams, A., Miceli, M., & Gebru, T. (2022). The Exploited Labor Behind Artificial Intelligence. Noema Magazine.

World Economic Forum (May 2022), AI Procurement in a box, https://www.weforum.org/reports/ai-procurement-in-a-box/

TRILATERAL RESEARCH
Ethical AI

UKRI Innovate UK

BridgeAI