RAND EUROPE

Examining the landscape of tools for trustworthy AI in the UK and the US

Current trends, future possibilities, and potential avenues for collaboration

Salil Gunashekar, Henri van Soest, Michelle Qu, Chryssa Politi, Maria Chiara Aquilino and Gregory Smith

For more information on this publication, visit www.rand.org/t/RRA3194-1

About RAND Europe
RAND Europe is a not-for-profit research organisation that helps improve policy and decision making through research and analysis. To learn more about RAND Europe, visit www.randeurope.org.

Research Integrity
Our mission to help improve policy and decision making through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behaviour. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/principles.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK

Cover: Adobe Stock

Preface

Over the years, there has been a proliferation of frameworks, declarations and principles from various organisations around the globe to guide the development of trustworthy artificial intelligence (AI). These frameworks articulate the foundations for the desirable outcomes and objectives of trustworthy AI systems, such as safety, fairness, transparency, accountability and privacy. However, they do not provide specific guidance on how to achieve these objectives, outcomes and requirements in practice. This is where ***tools for trustworthy AI*** become important. Broadly, these tools encompass specific methods, techniques, mechanisms and practices that can help to measure, evaluate, communicate, improve and enhance the trustworthiness of AI systems and applications.

Against the backdrop of a fast-moving and increasingly complex global AI ecosystem, this study mapped UK and US examples of developing, deploying and using tools for trustworthy AI. The research also identified some of the challenges and opportunities for UK–US alignment and collaboration on the topic and proposes a set of practical priority actions for further consideration by policymakers. The report's evidence aims to inform aspects of future bilateral cooperation between the UK and the US governments in relation to tools for trustworthy AI. Our analysis also intends to stimulate further debate and discussion among stakeholders as the capabilities and applications of AI continue to grow and the need for trustworthy AI becomes even more critical.

This rapid scoping study was conducted between November 2023 and January 2024 and was commissioned by the British Embassy Washington via the UK Foreign, Commonwealth and Development Office (FCDO) and the UK Department for Science, Innovation and Technology (DSIT). We would like to thank the project team at the British Embassy Washington for their support and guidance throughout the study. We are grateful for their valuable feedback and constructive guidance. In particular, we would like to thank Joe Cowen, Deepa Mani, Jonathan Tan and Alyssa Hanou. We would also like to thank our quality assurance reviewers at RAND Europe, Erik Silfversten and Sana Zakaria, for their feedback on drafts of the report. Finally, we are very grateful to the stakeholders who kindly agreed to participate in the interviews and crowdsourcing exercise.

RAND Europe is a not-for-profit research organisation that aims to improve policy and decision making in the public interest, through research and analysis. RAND Europe's clients include European governments, institutions, non-governmental organisations and firms with a need for rigorous, independent, multidisciplinary analysis.

The findings and analysis within this report represent the views of the authors and are not official government policy. For more information about RAND Europe or this document, please contact:

Salil Gunashekar (Deputy Director, Science and Emerging Technology Research Group)
RAND Europe
Eastbrook House, Shaftesbury Road
Cambridge CB2 8DR
United Kingdom

Email: sgunashe@randeurope.org

Henri van Soest (Senior Analyst, Defence and Security Research Group)
RAND Europe
Rue de la Loi 82 / Bte 3
1040 Brussels
Belgium

Email: vansoest@randeurope.org

Executive summary

Background and context

The pace of progress of AI has been rapid in recent years. AI is already being used in many fields and is a technology that could bring significant benefits to society, such as enhancing productivity, innovation, health, education and well-being. However, AI and its progress also pose major risks and challenges – including social, ethical, legal, economic and technical – that need to be addressed to ensure that AI is trustworthy. Consequently, AI has become a critical area of interest for stakeholders around the globe and there have been many discussions and initiatives to ensure that AI is developed and deployed in a responsible and ethical manner.

In general, AI systems and applications are regarded as trustworthy when they can be reliably developed and deployed without adverse consequences to individuals, groups or society.

While there is no universally accepted definition of the term trustworthy AI, various stakeholders – governments and international organisations alike – have proposed their own definitions, which characterise trustworthy AI based on a series of principles or guidelines that often overlap across definitions. These include such characteristics as fairness, transparency, accountability, privacy, safety and explainability.

Tools for trustworthy AI are specific approaches or methods to help make AI more trustworthy and can help to bridge the gap between the high-level AI principles and characteristics, on the one hand, and the practical implementation of trustworthy AI, on the other.

These tools encompass methods, techniques, mechanisms and practices that can help to measure, evaluate, communicate, improve and enhance the trustworthiness of AI systems. Thus, the goal of tools for trustworthy AI is to provide developers, policymakers and other stakeholders with the resources they need to ensure that AI is developed and deployed in a responsible and ethical manner. In Chapter 1 and Annex A, we provide more information about what we mean by trustworthy AI and tools for trustworthy AI in the context of this study.

Study objectives and research approach

The aim of this study was to examine the range of tools designed for the development, deployment and use of trustworthy AI in the United Kingdom and the United States.[1] The study identified challenges, opportunities and considerations for policymakers for future UK–US alignment and collaboration on tools for trustworthy AI. The research was commissioned by the British Embassy Washington, via the FCDO and DSIT. The study was conducted over eight weeks, between November 2023 and January 2024.

We used a mixed-methods approach to carry out the research. This involved a focused scan and review of documents and databases to identify examples of tools for trustworthy AI that have been developed and deployed in the UK and the US. We carried out interviews with experts connected to some of the identified tools and with wider stakeholders with understanding of tools for trustworthy AI. In parallel, we also conducted an online crowdsourcing exercise with a range of experts to collect additional information on selected examples of tools. Further details about the methodology are provided in Chapter 1 and Annex B.

Overview of the landscape of tools for trustworthy AI in the UK and the US

In the box below, we provide a descriptive overview of the range of tools identified that considers these tools' characteristics, similarities and differences and how these tools are being used in practice in the UK and the US. Further details about each key finding below are provided in Chapter 2.

1 In this study, we characterised trustworthy AI based on the fundamental underlying principles and/or characteristics of AI proposed by four major stakeholders across the world – specifically, the UK, the US, the European Commission and the Organisation for Economic Co-operation and Development. In Chapter 1 and Annex A, we provide further details about these principles and characteristics.

Box 1: Overview of the UK and the US landscapes of tools for trustworthy AI

Indicative of a potentially fragmented landscape, we identified 233 tools for trustworthy AI, of which roughly 70% (n=163) were associated with the US, 28% (n=66) were associated with the UK, and the remainder (n=4) represented a collaboration between US and UK organisations. Broadly, the tools can be categorised as technical, procedural or educational (drawing on the classification used by the Organisation for Economic Co-operation and Development), which further encompass a range of characteristics and dimensions associated with trustworthy AI.

The landscape of tools for trustworthy AI in the US is more technical in nature, while the landscape in the UK is observed to be more procedural. Roughly 72% (n=119) of the US tools were technical in nature, while 56% (n=37) of the UK tools were technical in nature. 30% (n=49) of the US tools were procedural, compared with 58% (n=38) of the UK tools. Finally, 9% (n=16) of the US tools were educational, compared with 12% (n=8) of the UK tools.

Compared to the UK, the US has a greater degree of involvement of academia in the development of tools for trustworthy AI. Roughly 27% (n=45) of the US tools were developed by academia or collaboratively between academia and external partners, such as industry or non-profit organisations. By contrast, 9% (n=6) of the UK tools for trustworthy AI involved academia.

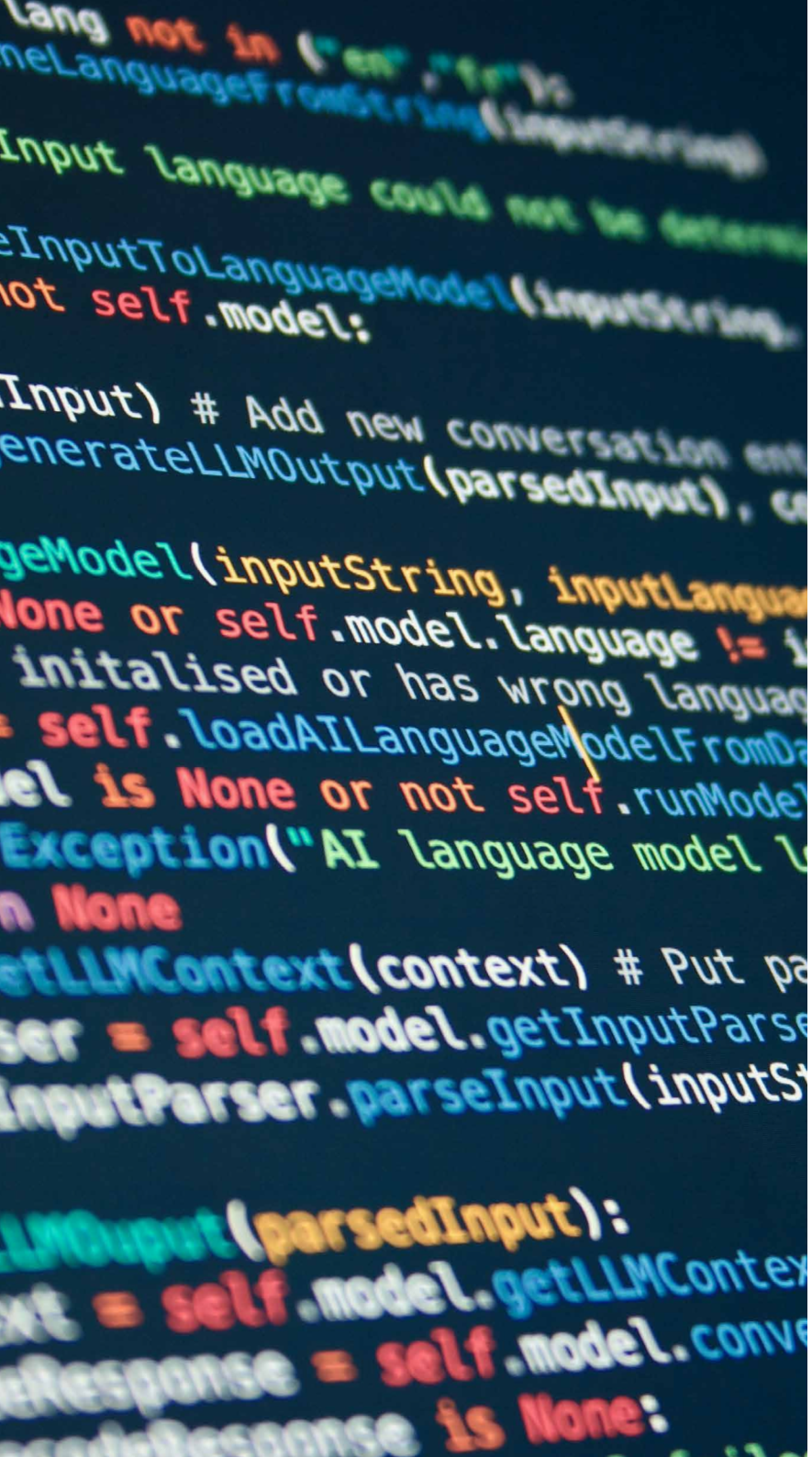
Large US technology companies are developing wide-ranging toolkits to make AI products and services more trustworthy.

There is limited evidence about the formal assessment of tools for trustworthy AI.

Some non-AI companies are developing their own internal guidelines on AI trustworthiness to ensure they comply with ethical principles.

The development of multimodal foundation models has increased the complexity of developing tools for trustworthy AI.

Source: RAND Europe analysis.



Proposed considerations for policymakers

We propose a series of considerations for stakeholders – primarily policymakers – involved in the tools for trustworthy AI ecosystem in the UK and the US (see Figure 1). Developing and using tools for trustworthy AI are not sufficient actions by themselves.

The tools need to be complemented by a collaborative and inclusive approach that involves multiple perspectives and actors, such as governments, businesses, civil society, academia and international organisations.

We offer these suggestions as a set of cross-cutting practical actions. When taken together and combined with other activities and partnership frameworks – for example, the Atlantic Declaration[2] – in the wider context of AI regulatory policy debates and collaboration, these actions could potentially help contribute to a more linked-up, aligned and agile ecosystem between the UK and the US. We provide further details about each proposed action in Chapter 3.

2 DBT et al. (2023).

Figure 1. Practical considerations for UK and US policymakers to help build a linked-up, aligned and agile ecosystem

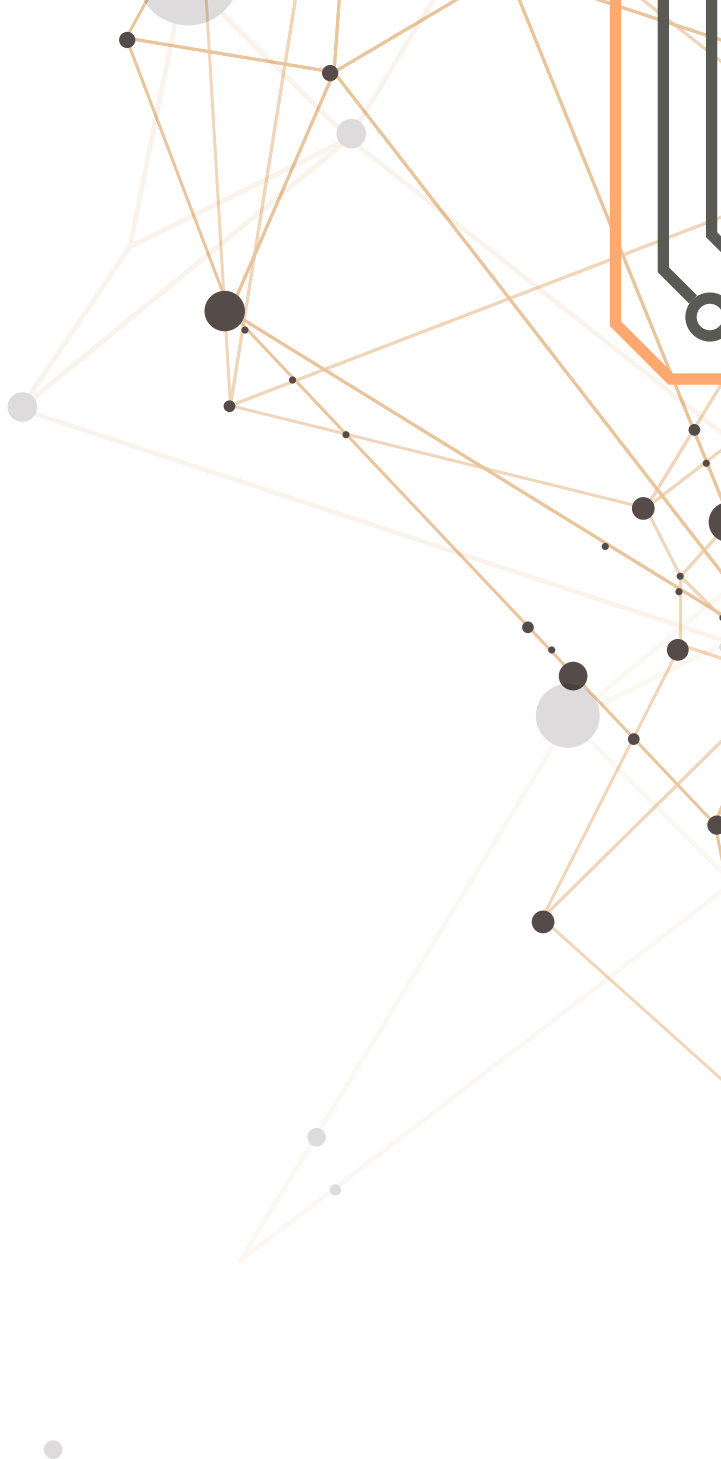
Potential stakeholders to involve across the different actions: Department for Science, Innovation and Technology (including the Responsible Technology Adoption Unit and UK AI Safety Institute); Foreign, Commonwealth & Development Office (including the British Embassy Washington); AI Standards Hub; UK Research and Innovation; AI Research Resource; techUK; Evaluation Task Force in the UK; Government Office for Science; National Institute of Standards and Technology; US AI Safety Institute; National Science Foundation; National Artificial Intelligence Research Resource; US national laboratories; Organisation for Economic Co-operation and Development; European Commission; United Nations (and associated agencies); standards development organisations.

Source: RAND Europe analysis

Table of contents

Abbreviations and acronyms

AI	Artificial intelligence
AIRR	AI Research Resource (UK)
AISI, UK	AI Safety Institute (UK)
AISI, US	AI Safety Institute (US)
ANSI	American National Standards Institute
BSI	British Standards Institution
CDEI	Centre for Data Ethics and Innovation (UK)
EC	European Commission
EU	European Union
EU–US TTC	EU–US Trade and Technology Council (EC)
FCDO	Foreign, Commonwealth and Development Office (UK)
GPAI	Global Partnership on Artificial Intelligence
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
NAIRR	National Artificial Intelligence Research Resource (US)
NIST	National Institute of Standards and Technology (US)
NSF	National Science Foundation (US)
OECD	Organisation for Economic Co-operation and Development
RTA	Responsible Technology Adoption Unit (UK)
UNESCO	United Nations Educational, Scientific and Cultural Organization



Chapter 1

What is this study about?

1.1. Background and context

Across the world, artificial intelligence (AI) is rapidly transforming various aspects of our society, from healthcare and education to finance and entertainment. However, as AI becomes more capable and pervasive, it raises ethical, social and legal challenges that need to be continuously addressed as the technology advances at pace. Consequently, AI has become a crucial area of interest for stakeholders around the world. How safe, secure and reliable is an AI system? How can we ensure that AI systems are aligned with human values and respect human rights? How can we prevent and mitigate the potential harms of AI, such as bias, discrimination, manipulation and deception? How well and transparently are the decisions and actions of AI systems explained? How can we foster trust and confidence in AI among consumers and the public? These and other related questions have prompted much debate and discussion over the years about 'trustworthy AI' and how to ensure that AI systems and applications are trustworthy.

1.1.1. What do we mean by trustworthy AI and tools for trustworthy AI in the context of this study?

Trustworthy AI is a wide-ranging and complex concept. **In general, AI systems and applications are regarded as trustworthy when they can be reliably developed and deployed without adverse consequences to individuals, groups or society.** While there is no universally accepted definition of the term trustworthy AI, various stakeholders – governments and international organisations alike – have proposed their own definitions, which characterise trustworthy AI based on a series of principles or guidelines that often overlap across definitions. These include such characteristics as fairness, transparency, accountability, privacy, safety and explainability.

Over the years, discussions around trustworthy AI have prompted the development of various frameworks and principles for trustworthy AI, such as the European Commission's (EC) Ethics Guidelines for Trustworthy AI[3]; the Organisation for Economic Co-operation and Development (OECD) AI Principles[4]; the United Nations Educational, Scientific and Cultural Organization's (UNESCO) Recommendation on the Ethics of AI[5]; and, more recently, the underpinning principles of the UK government's AI regulation white paper,[6] the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,[7] and the White House Office of Science and Technology Policy's Blueprint for an AI Bill of Rights.[8] These frameworks and principles, to varying degrees of detail, lay the important foundations that outline the desirable outcomes and objectives of trustworthy AI systems – as well as the trustworthiness of the processes and involved stakeholders – throughout the system's life cycle. However, they do not provide specific guidance on how to achieve these objectives, outcomes and requirements in practice.

This is where tools for trustworthy AI become very relevant. **Tools for trustworthy AI are specific approaches or methods to help make AI more trustworthy and can help to bridge the gap between the high-level AI principles and characteristics, on the one hand, and the practical implementation of trustworthy AI, on the other.** Broadly, these tools encompass methods, techniques, mechanisms and practices that can help to measure, evaluate and communicate the trustworthiness of AI systems and applications (where trustworthiness can be characterised by different dimensions as listed above). They can also help to improve and enhance the trustworthiness of AI systems and applications by identifying and addressing potential issues and risks. Thus, the goal of tools for trustworthy AI is to provide developers, policymakers and other stakeholders with the resources they need to ensure that AI is developed and deployed in a responsible and ethical manner.

In this report, we focus on the state of play of tools for trustworthy AI in the UK and the US ecosystems. We characterised the trustworthiness of AI based on the fundamental underlying principles proposed by four major stakeholders in different regions across the world that are currently

3 EC (2019).
4 OECD (2019).
5 UNESCO (2021).
6 DSIT (2023).
7 The White House (2023a).
8 The White House (2022).

actively involved in key AI-related discussions and debates – specifically, the UK, the US, the EC and the OECD. In Table 1, we outline the key dimensions of trustworthy AI covered by each stakeholder. In Annex A, we provide further details on these principles and characteristics. We have deliberately relied on an inclusive and holistic interpretation of trustworthy AI. Such an expansive characterisation fed into our methodology to identify tools in the UK and the US and allowed us to capture a variety of examples of tools that have been designed and developed for trustworthy AI.

Table 1. Key underlying principles and characteristics of trustworthy AI, from different stakeholders, that were used in this study

UK government[9]	National Institute of Standards and Technology (US)[10]	European Commission[11]	Organisation for Economic Co-operation and Development[12]
Five principles: • Safety, security and robustness • Appropriate transparency and explainability • Fairness • Accountability and governance • Contestability and redress	Seven characteristics: • Valid and reliable • Safe • Secure and resilient • Accountable and transparent • Explainable and interpretable • Privacy-enhanced • Fair – with harmful bias managed	Three components: • Lawful • Ethical • Robust Four ethical principles: • Respect for human autonomy • Prevention of harm • Fairness • Explicability Seven requirements: • Human agency and oversight • Technical robustness and safety • Privacy and data governance • Transparency • Diversity, non-discrimination and fairness • Societal and environmental well-being • Accountability	Five principles: • Inclusive growth, sustainable development and well-being • Human-centred values and fairness • Transparency and explainability • Robustness, security, and safety • Accountability

Source: RAND Europe synthesis of the respective sources cited in the heading row

9 DSIT (2023, 2024a).
10 NIST (2023).
11 EC (2019).
12 OECD (2019).

1.2. Objectives of the study

Against the backdrop of a fast-moving and increasingly complex global AI ecosystem, the aim of the research was to examine the range of tools designed to help implement trustworthy AI systems and applications in the UK and the US. The research was commissioned by the British Embassy Washington via the UK Foreign, Commonwealth and Development Office (FCDO) and the UK Department for Science, Innovation and Technology (DSIT). Specifically, the research mapped UK and US examples of developing, deploying and using tools for trustworthy AI and, where relevant, industry uses of practical tools. The research also identified some of the challenges and opportunities for UK–US alignment and collaboration on the topic, and proposes a set of practical considerations for policymakers. The study was conducted over eight weeks, between November 2023 and January 2024.

The report's evidence aims to inform aspects of future bilateral cooperation between the UK and the US governments in relation to tools for trustworthy AI. The findings from this study are thus mainly targeted at policymakers. However, they are also likely to be of interest and relevance to other stakeholders involved in AI and wider technology policy, such as regulators, funders of research and innovation, and those working in academic and industry. Our analysis also intends to stimulate further debate and discussion among stakeholders as the capabilities and applications of AI continue to grow and the need for trustworthy AI becomes even more critical.

1.3. Overview of the methodology

We used a mixed-methods approach for the study. In the first phase of the research, we carried out scoping consultations (by interview and email) with three AI experts within the RAND Corporation who have knowledge of developments within the wider AI ecosystem, to help develop a better understanding of the state of play and to help identify key stakeholders to speak to in the next phase of the research.

In the second phase of the study, we carried out a focused scan and review of documents and databases to identify examples of tools for trustworthy AI that have been developed and deployed in the UK and the US. We created a database to capture the different examples and various types of information associated with them (including, for example, a short description of the tool; the developer of the tool; the country the tool was developed in; the timeline of development; the tool's objective; the tool type; the stage of development; and the sector(s) that the tool targeted). This enabled us to cross-analyse the tools to extract common themes and trends associated with the data, as well as notable divergences. In parallel, we also conducted an online crowdsourcing exercise with a range of experts to collect additional information on examples of tools and other material, such as relevant reports, articles and websites.[13] Alongside the desk research, we conducted ten semi-structured interviews with experts connected to some of the identified tools, as well as wider stakeholders across academia, industry, government and the third sector with understanding of tools for trustworthy AI.[14]

13 We reached out to 64 experts, based in the US, the UK and the EU.

14 The evidence from the interviews has been anonymised and cited throughout the report using unique interviewee identifiers (INT01, INT02, etc.).

In the final phase of the research, we cross-analysed the findings from the desk research – i.e. the longlist of tools identified – and complemented this analysis with information from the interviews. The resulting findings form the basis of the narrative and key takeaways presented in this report.

We provide more details about the research methodology and associated caveats in Annex B.

Chapter 2

What does the landscape of tools for trustworthy AI look like in the UK and the US?

In this chapter, we discuss what the landscape of tools for trustworthy AI looks like in the UK and the US, based on a cross-analysis of the document and database review and interviews. The chapter begins with a high-level descriptive overview of the range of tools identified, followed by an analysis on how these tools are being used in the context of trustworthy AI. Throughout the chapter, we highlight examples of AI tools to illustrate specific findings from the research.[15]

15 The examples we include in this report do not represent an endorsement of the tools or techniques or of the organisation developing them.

Box 2: Overview of the UK and the US landscapes of tools for trustworthy AI

Indicative of a potentially fragmented landscape, we identified 233 tools for trustworthy AI, of which roughly 70% (n=163) were associated with the US, 28% (n=66) were associated with the UK, and the remainder (n=4) represented a collaboration between US and UK organisations. Broadly, the tools can be categorised as technical, procedural or educational (drawing on the classification used by the Organisation for Economic Co-operation and Development), which further encompass a range of characteristics and dimensions associated with trustworthy AI.

The landscape of tools for trustworthy AI in the US is more technical in nature, while the landscape in the UK is observed to be more procedural. Roughly 72% (n=119) of the US tools were technical in nature, while 56% (n=37) of the UK tools were technical in nature. 30% (n=49) of the US tools were procedural, compared with 58% (n=38) of the UK tools. Finally, 9% (n=16) of the US tools were educational, compared with 12% (n=8) of the UK tools.

Compared to the UK, the US has a greater degree of involvement of academia in the development of tools for trustworthy AI. Roughly 27% (n=45) of the US tools were developed by academia or collaboratively between academia and external partners, such as industry or non-profit organisations. By contrast, 9% (n=6) of the UK tools for trustworthy AI involved academia.

Large US technology companies are developing wide-ranging toolkits to make AI products and services more trustworthy.

There is limited evidence about the formal assessment of tools for trustworthy AI.

Some non-AI companies are developing their own internal guidelines on AI trustworthiness to ensure they comply with ethical principles.

The development of multimodal foundation models has increased the complexity of developing tools for trustworthy AI.

Source: RAND Europe analysis

2.1. Overview of tools identified

The main data collection component of the research involved the collation of a range of tools for trustworthy AI in the UK and the US. This section presents a descriptive overview of the tools we identified – based on the data we had collated by mid-January 2024 – considering their characteristics, similarities and differences.[16] Further details about specific examples of tools are provided in Annex C and the accompanying Excel file containing the longlist of tools we identified.[17]

- Indicative of a potentially fragmented landscape, we identified a total of **233 tools for trustworthy AI spanning the US and the UK** (at the time of writing the report).[18] Some of the tools cover multiple jurisdictions across the globe: 163 (approximately 70%) tools are associated with the US; 66 (approximately 28%) are associated with the UK; and 4 tools represent a form of collaboration between US and UK organisations.
- The tools encompass a **wide range of characteristics and dimensions associated with trustworthy AI.** We organised these characteristics of trustworthy AI into the following categories: accountability, fairness; human well-being; performance; transparency; privacy and data governance; reskilling or upskilling; respect for human rights; robustness and digital society; safety; sustainability; and transparency and explainability. A total of 93 tools were linked to fairness, while 74 were linked with accountability. However, many of the tools address several dimensions of trustworthy AI, and there is overlap between the different dimensions. Other dimensions, such as sustainability (n=16), are less prevalent.
- Some tools are **technical** in nature, which means they try to offer solutions in the form of code or algorithms that can be run on AI models or datasets to ensure the trustworthiness of AI systems. Other tools are **procedural** in nature, which means they offer compliance-based solutions where AI models are evaluated and red-teamed to discern their trustworthiness. Other tools are **educational** and aim to make specific stakeholders or the wider public aware of trustworthy AI. In our dataset, we found that 155 tools were technical in nature, 87 tools were procedural, and 19 tools were educational. We found differences in the distribution of these categories in the US and the UK ecosystems. Of the US tools for trustworthy AI identified, 119 tools (approximately 72%) were technical in nature. In the UK, by contrast, 37 tools (approximately 56%) were technical in nature.[19] By comparison, the US had 49 procedural tools (approximately 30% of all US tools), while the UK had 38 procedural tools (approximately 58% of all UK tools). Finally, the US had 16 educational tools (approximately 9% of all US tools), while the UK had 8 educational tools (approximately 12% of all UK tools).

16 The categories we used in this analysis align with those used in the OECD Catalogue of Tools & Metrics for Trustworthy AI (OECD 2021).

17 The Excel spreadsheet was populated based on the information contained in the source data we consulted or using our best understanding of the information associated with the tool that we analysed. We recognise that some of the information contained in the source data may not be the most up-to-date information linked to that tool. Furthermore, it is possible for a tool to be linked to more than one category. For example, a tool may be classified as both technical and educational in nature. As a result, the sum of these classification values may be larger than the number of tools identified.

18 It is worth noting that the total figure reported here reflects each tool example we identified in the underpinning source data – this includes an aggregation of individual tools as well as toolkits (that may, in some examples, include constituent tools).

19 Although this cannot be verified without further in-depth examination of the different tools, the fact that procedural tools are less prominent in the US may be linked to cultural differences and a relatively more general lack of support for certification compared with technical solutions in the US context (INT10).

- Across the three broad categories, the tools encompass **different specific tool types.** There is a diverse range of tool types, such as audit processes, checklists, guidelines, standards and sectoral codes of conduct. For example, we identified 138 toolkits or software solutions, of which 115 were from the US (83%) and 23 were from the UK (17%).[20] We identified 18 audit processes, of which 5 were from the US (32%) and 13 from the UK (68%).
- The tools also had **different levels of maturity.** Using the OECD grouping for tool readiness,[21] we identified tools that were: under development; presented in a published document; in the product stage; or implemented in multiple projects. For example, we found 75 tools that were under development and 111 tools that have been implemented in multiple projects.
- The tools were **developed by a range of stakeholders** across diverse types of organisations spanning industry, academia and not-for-profit organisations. There was also **collaboration** between these categories. For example, Microsoft Research separately worked with the University of Pennsylvania[22]; the University of Washington[23]; and the Montreal AI Ethics Institute, McGill University (both in Quebec, Canada) and Carnegie Mellon University (also in Pennsylvania).[24] Google worked with the Courant Institute of Mathematical Sciences at New York University.[25] However, we found differences between the involvement of academia in the US and the UK tools for trustworthy AI ecosystems. Of the 163 US tools we identified, 45 tools (approximately 27%) were developed by academia or collaboratively between academia and external partners, such as industry or non-profit organisations. By contrast, of the 66 UK tools we identified, 6 were developed by academia (approximately 9%), and we only found 1 example of a British academic institution working together with external partners.[26]

---

20 While toolkits and software could be seen as distinctive tool types, the OECD catalogue combines them into a single category. We decided to maintain this category for the purposes of this study.

21 OECD (2024a).

22 Kearns et al. (2018).

23 Covert et al. (2020).

24 Gupta et al. (2020).

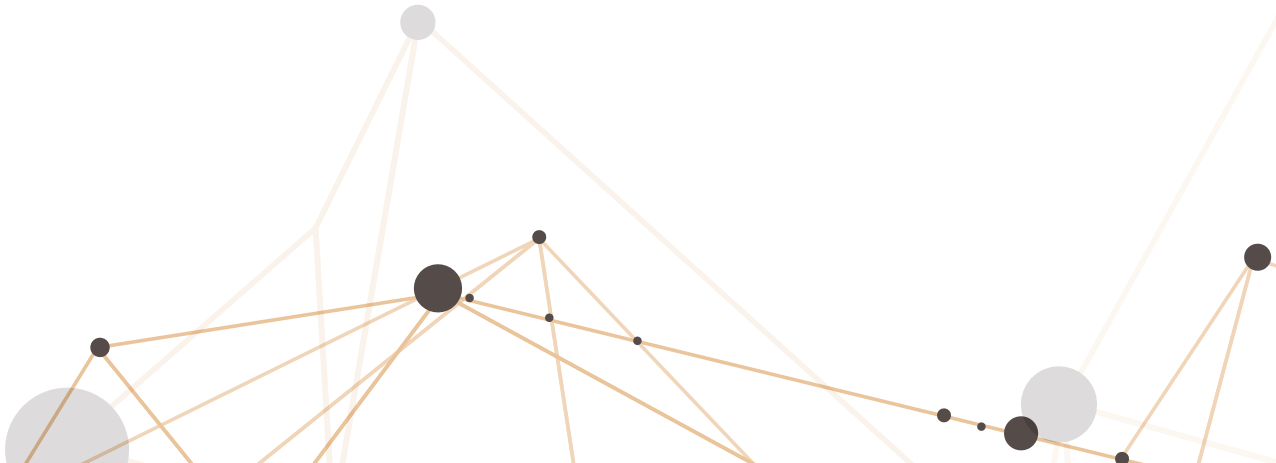
25 Cortes et al. (2017).

26 Berditchevskaia et al. (2021).

2.2. The landscape of trustworthy AI in the UK and the US is moving from principles to practice, and high-level guidelines are increasingly being complemented by more specific, practical tools

The landscape of trustworthy AI in the UK and the US is complex and multifaceted. Both the technical advancement of AI and attempts to make AI more trustworthy are under development, and the respective ecosystems are changing rapidly. Interviewees pointed out that the need to make AI trustworthy has moved through stages of development over time, which is also reflected in the wider landscape of tools.[27] Initially, thinking about the ethical implications of AI led to the development of high-level guidelines, such as those discussed in Section 1.1. Over time, these ethical principles started getting translated into attempts to try to regulate AI, for example, the EC's proposed legal framework on AI.[28] The need to operationalise these proposed regulations in turn led to a discussion about AI standardisation.[29]

We found that the examples of tools for trustworthy AI we identified broadly reflect these developments. There are several high-level guidelines that set out principles around AI trustworthiness that are also regarded as 'tools' for trustworthy AI. As noted in Chapter 1, while these guidelines are helpful in furthering the development of trustworthy AI as a concept, they cannot directly be operationalised and applied in practice. There is also a category of tools that consist of more specific technical tools. These are developed by and targeted at software engineers and developers. In addition, compliance-based approaches are being developed that can help non-specialised businesses to evaluate the deployment of AI models.[30]



27 INT07; INT09.
28 EC (2024a); INT02.
29 INT07; INT08; INT10.
30 INT09.

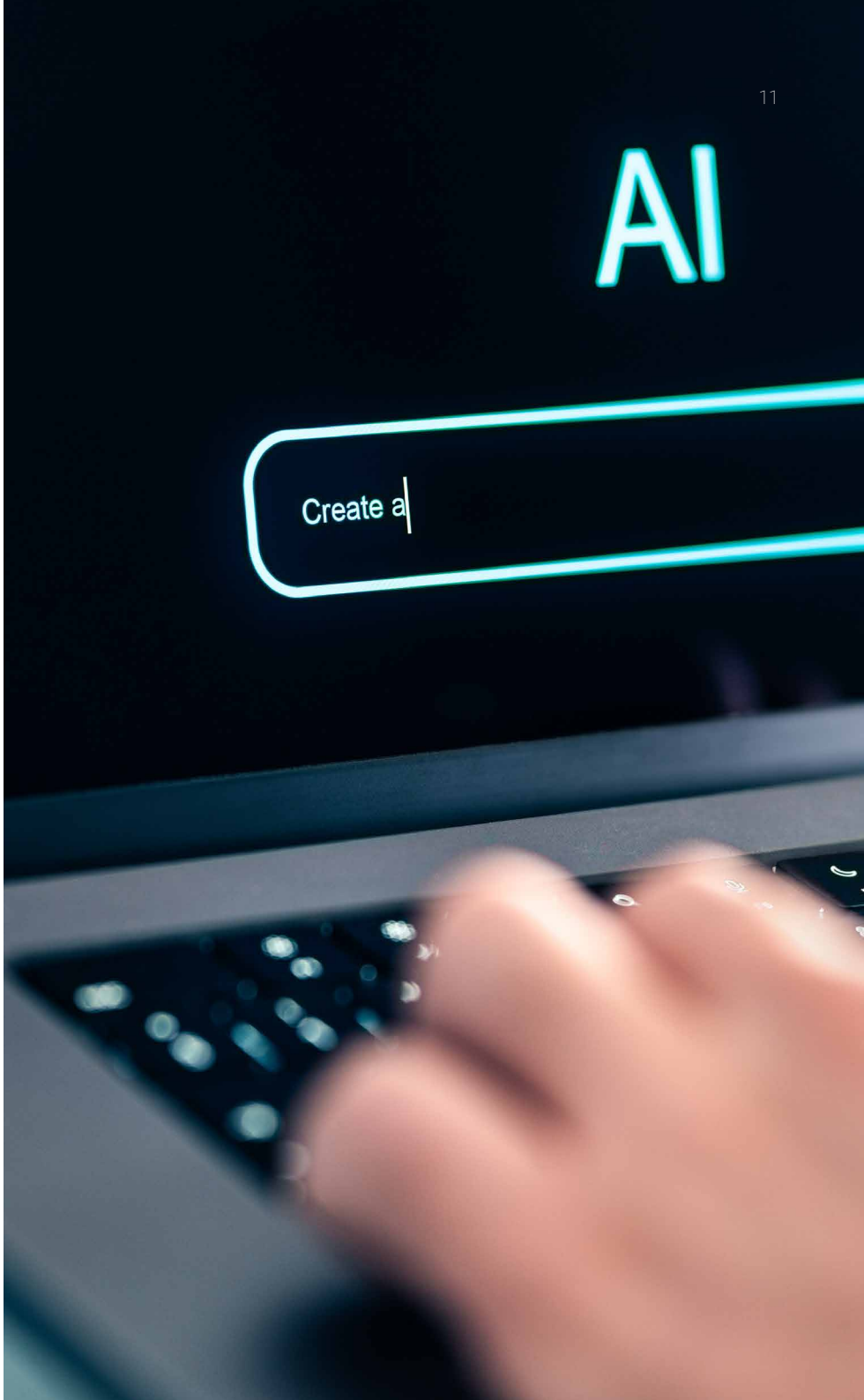
Box 3: AI4People Ethical Framework for a Good AI Society

AI4People is an international consortium of researchers. Its Ethical Framework for a Good AI Society outlines the risks and opportunities of a widespread implementation of AI. It also sets out the principles that are key for ensuring that AI has a positive impact on society. The principle of 'beneficence' covers the promotion of well-being, the preservation of dignity, and environmental sustainability. 'Non-maleficence' covers privacy and security requirements. 'Autonomy' is the balance between delegating decision making to AI and retaining human decision-making power. 'Justice' covers the need to preserve prosperity and solidarity. 'Explicability' means that humans must be able to understand AI decision making.[31]

Box 4: Microsoft Counterfit

Microsoft has developed Counterfit, a command-line tool that provides a generic automation layer that can be applied to AI models to assess their security. It uses a range of adversarial attack models that can be used to red-team AI models. It uses a similar workflow and setup to other popular offensive tools used by cybersecurity professionals. This technical tool aims to help software engineers improve the security of their AI models by allowing them to discover vulnerabilities before they are exploited.[32]

31 Floridi et al. (2019).
32 Kumar (2021).



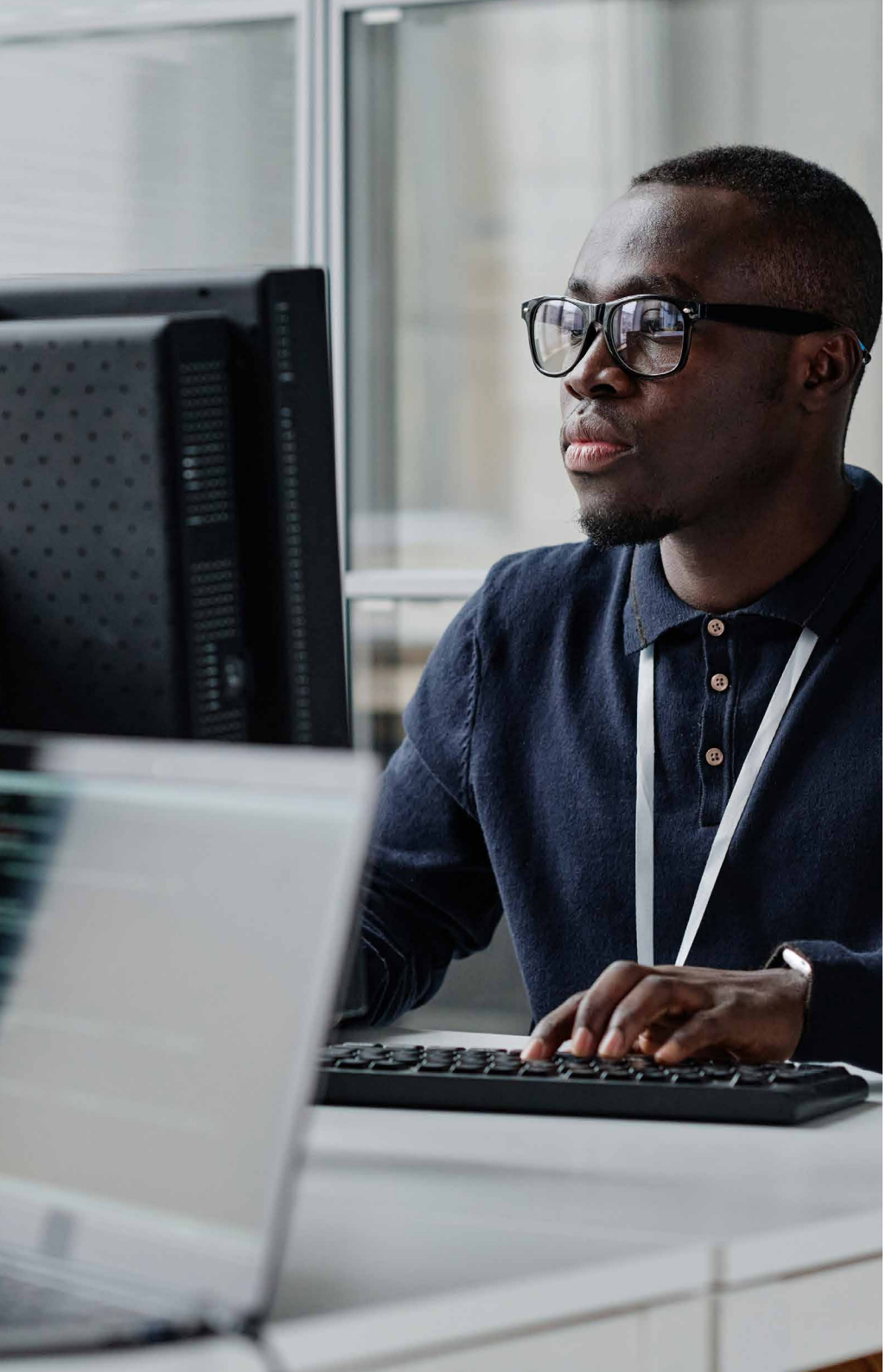
2.3. Large US technology companies are developing wide-ranging toolkits to make AI products and services more trustworthy

Several major technology companies, such as IBM, Microsoft and Google, have developed toolkits that bring together a collection of separate tools to address various aspects of trustworthy AI. These companies have large research divisions that are funded through the company's commercial activities. Because these companies are often developing AI models themselves, they have a 'head start' in understanding the functionality of the models, which can help them in developing tools for making the models more trustworthy. Researching how the various aspects of AI trustworthiness can be measured and improved can therefore form an integral part of the company's product development activities.[33]

The toolkits are often meant to be general 'wrappers' that contain several individual, specific approaches and that often contain constituent tools and metrics.[34] For example, the approaches in the toolkit can be applied to the systems of users in a 'pick-and-mix' fashion, and they can be plugged into a model to run different tests and create a 'model report card'.[35] Additionally, the toolkits can help with a range of other issues, such as finding bugs in the data input or issues with model weights.[36]

It was noted that the toolkits developed by the large technology companies appear to be regularly updated and are often publicly available.[37] However, the companies also work together with specific clients to build out toolkits for their own specific purposes and in their own context. These clients will often have complicated and unique system architectures in place, which require a tailored solution.[38] Alternatively, specific clients may have highly specific AI applications that require a different emphasis in terms of testing and evaluation.[39] This may be the case, for example, for military AI-enabled targeting systems. In our research, we primarily found examples of large US-based technology companies.

33 INT05.
34 Tools are approaches to analyse or improve the trustworthiness of an AI model, while metrics are mathematical formulas for measuring certain technical requirements relating to trustworthy AI.
35 INT05.
36 INT05.
37 INT05; INT06.
38 INT05.
39 INT05.



Box 5: IBM Toolkits

IBM has developed a range of toolkits that bring together multiple tools and metrics, including:

- IBM Research AI Fairness 360: This toolkit includes metrics for testing biases, explanations and instructions for using these metrics, and algorithms for mitigating bias in datasets and models.[40]
- IBM Research AI Privacy 360: This toolkit includes tools to support the assessment of privacy risks in AI applications and to enable these AI applications to comply with privacy regulations.[41]
- IBM Research AI Explainability 360: This toolkit includes metrics across the spectrum of explainability.[42]
- IBM Research Uncertainty Quantification 360: This toolkit includes metrics and algorithms that help in the estimation, evaluation and communication of uncertainty in AI and that can help in reducing uncertainty.[43]
- IBM Research AI FactSheets 360: This toolkit contains factsheets that outline different aspects of AI governance.[44]
- IBM Adversarial Robustness 360: This toolkit contains tools that can help in the evaluation and defence of AI applications against adversarial threats, such as evasion, poisoning and extraction of data.[45]

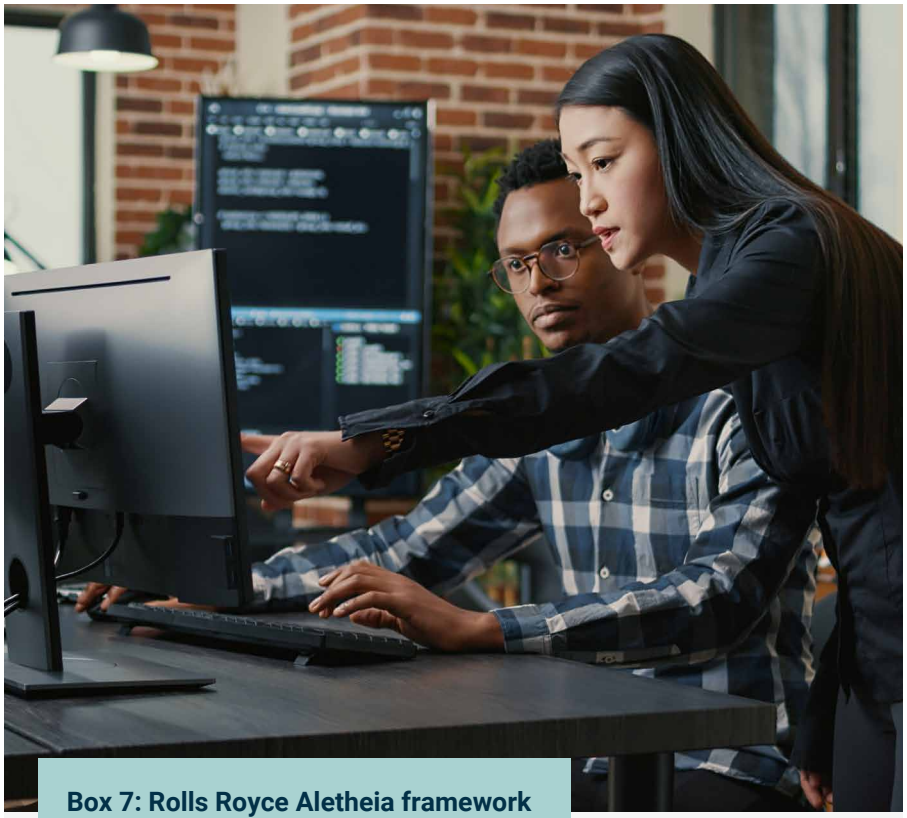
40 IBM Research (2024a).
IBM Research (2024b).
IBM Research (2024c).
IBM Research (2024d).
IBM Research (2024e).
IBM Research (2024f).

2.4. Some non-AI companies are developing their own internal guidelines on AI trustworthiness to ensure they comply with ethical principles

There are some large companies that are not explicitly AI companies or large 'tech' companies. These include, for example, telecommunications companies and industrial companies that engage with and use advanced data science and machine learning techniques. While the large technology companies developing AI tools take a developer view, the non-AI companies tend to take a practical approach rooted in their current working processes. This practical grounding makes these tools particularly interesting for assessing the potential practical impacts of tools for trustworthy AI. We found examples of both UK and US non-AI companies developing similar internal guidelines.

Box 6: Comcast's Project Guardrail

US telecommunications company Comcast has developed a set of security and privacy requirements for AI applications that serve as guardrails against outputs or uses of the AI model that cannot be considered trustworthy. These requirements consists of a baseline that all AI applications developed and deployed within Comcast have to meet, as well as two additional sets that are specific to continuously learning models and user-interacting models.[46]



Box 7: Rolls Royce Aletheia framework

Rolls Royce has developed the Aletheia framework, which is a framework to govern the ethical and responsible use of AI. It consists of a toolkit that addresses 32 facets of social impact, trust, transparency and governance. The goal of the framework is to guide developers, executives and boards on the deployment of AI. The toolkit was first developed for internal use by Rolls Royce, and the company then decided to make it public.[47]

46 Comcast (2023).
47 Rolls Royce (2023).

2.5. There is limited evidence about the formal assessment of tools for trustworthy AI

Based on the evidence we have reviewed in this study, tools for trustworthy AI do not appear to be formally assessed very often for their quality or effectiveness (e.g. post-deployment), and there is limited publicly available evidence of stakeholders sharing their experiences of developing, deploying or using these tools in practice.[48] Moreover, there do not appear to be any systematic evaluations across portfolios of tools covering similar approaches to extract key learnings, for example, in relation to barriers, enablers and good practices.[49] The OECD, through its online Catalogue of Tools & Metrics for Trustworthy AI, has started to share experiences of stakeholders that have used tools, through a series of use cases.[50] Similarly, DSIT's online Portfolio of AI Assurance Techniques features a growing number of examples of AI assurance techniques that have been used in practice to measure, evaluate and communicate the trustworthiness of AI systems across a range of real-world use cases.[51]

Box 8: Institute for the Future of Work's investigation of AI auditing tools

One of the few examples we found of an assessment of tools for trustworthy AI was an investigation of AI auditing tools conducted by the Institute for the Future Work (IFOW), a UK-based research and development institute. The use of AI applications in recruitment in hiring and recruitment can lead to inequality, bias and discrimination in decision making. AI auditing tools are often touted as a solution. However, the IFOW found that many of these tools are not robust enough to ensure compliance with UK Equality Law, good governance and best practice. For example, existing AI auditing tools typically provide only a snapshot assessment of bias in an AI system, whereas the effects of bias should be considered over time. Overall, the IFOW found that a mere mechanistic application of technical AI auditing tools would be insufficient to safeguard equality in hiring and recruiting practices.[52]

48 INT03; INT10.
49 INT01; INT05.
50 OECD (2024b).
51 CDEI and DSIT (2024a).
52 Graham et al. (2020).

2.6. The development of multimodal foundation models has increased the complexity of developing tools for trustworthy AI

Several interviewees highlighted that the development of multimodal foundation models has increased the complexity of the challenges involved in ensuring that AI is trustworthy.[53] While there are tools for individual applications, such as text generation and image recognition, the combination of different modules into a model makes the behaviour of these models significantly more complex.[54] Furthermore, the current speed of development means that existing models are quickly surpassed by newer models. This pace of development also puts 'pressure' on the speed of model evaluations, because if the evaluation takes too long, it could lose its purpose.[55] It was noted that potentially different foundation models will start to look quite 'similar' as the existing data becomes exhausted and several foundation models eventually get trained on the same data.[56] Researchers could develop general principles and frameworks that are model-agnostic that can be applied regardless of the model that is being evaluated.[57]

Box 9: The Partnership on AI's Guidance for Safe Foundation Model Deployment

The Partnership on AI has developed Guidance for Safe Foundation Model Deployment. The partnership will send foundation model providers tailored guidance in the form of a set of good practices to be followed throughout the deployment process, tailored to the specific model and its release modalities. The Guidance for Safe Foundation Model Deployment is meant to be a living document that can be updated in response to the further development of foundation model capabilities.[58]

53 INT04; INT05; INT06.
54 INT04; INT05.
55 INT05.
56 INT04; INT05.
57 INT05.
58 PAI (2023).

Chapter 3

What actions should be considered looking ahead?

In this chapter, we offer some concluding remarks on the study's findings. We reflect on the potential future direction of the tools for trustworthy AI ecosystems, focusing on the landscapes in the UK and the US, but also considering some of the notable wider policy developments that are taking place across the globe. We propose a series of priority considerations for stakeholders – primarily policymakers in the UK and the US – involved in developing the trustworthy AI ecosystem.

Box 10: Key takeaways from this chapter

Practical considerations for policymakers

Action 1: Link up with relevant stakeholders to proactively track and analyse the landscape of tools for trustworthy AI in the UK, the US and beyond.

Action 2: Systematically capture experiences and lessons learnt on tools for trustworthy AI, share those insights with stakeholders, and use them to anticipate potential future directions.

Action 3: Promote the consistent use of a common vocabulary for trustworthy AI among stakeholders in the UK and the US.

Action 4: Encourage the inclusion of assessment processes in the development and use of tools for trustworthy AI to gain a better understanding of their effectiveness.

Action 5: Continue to partner and build diverse coalitions with international organisations and initiatives, and to promote interoperable tools for trustworthy AI.

Action 6: Join forces to provide resources such as data and computing power to support and democratise the development of tools for trustworthy AI.

Source: RAND Europe analysis

3.1. Practical considerations for policymakers

AI is a technology that can bring significant benefits to society, such as enhancing productivity, innovation, health, education and well-being. However, AI also poses considerable risks and challenges, such as social, ethical, legal, economic and technical issues, that need to be addressed to ensure that AI is trustworthy, responsible and human-centric. One of the key aspects of trustworthy AI is the availability and use of specific tools that can help stakeholders in the AI ecosystem, such as developers, deployers, users, regulators and policymakers, to design, implement, monitor and evaluate AI systems and applications in alignment with the principles of trustworthy AI.

This study provides a high-level analysis of the landscape of tools for trustworthy AI in the UK and the US. As we have demonstrated, these tools can vary greatly and are being widely developed by different stakeholders in the ecosystem across industry, academia and government. The tools range from software programmes to procedural guidelines and standards, from educational initiatives and training to certifications and quality marks, and from documentation and reporting to auditing and oversight.

Reflecting on our analysis, we propose a series of practical considerations for stakeholders involved in the trustworthy AI ecosystem. These actions are not intended to be definitive or exhaustive; rather, they serve as a set of topics for further discussion and debate by relevant policymakers and, more generally, by stakeholders in the AI community associated with and interested in trustworthy AI. The actions we have proposed are wide ranging and relate to complex issues associated with AI trustworthiness and broader AI oversight-related matters, and they will require multiple stakeholders in the UK, the US, and beyond to take the initiative and work together in a coordinated manner. The suggested actions, along with the data collected through them, could potentially help further inform and support the framing of a robust consensus on tools for trustworthy AI, which could be particularly helpful for future discussions about wider AI oversight.

Developing and using tools for trustworthy AI are not sufficient actions by themselves.

The tools need to be complemented by a collaborative and inclusive approach that involves multiple perspectives and actors, such as governments, businesses, civil society, academia and international organisations.

We therefore offer the following suggestions as a set of cross-cutting practical actions that, when taken together and combined with other activities and partnership frameworks (for example, the Atlantic Declaration)[59] in the wider context of emerging AI regulatory policy debates and collaboration, could potentially help contribute to a more linked-up, aligned and agile trustworthy AI ecosystem between the UK and the US. We also suggest key stakeholders who potentially could be involved in developing and implementing some of the proposed actions. Where relevant, in the narrative accompanying some of the actions, we have specified the high-level role that some of the notable stakeholders might take on – based on our current understanding of the remit of those stakeholders. It is beyond the scope of this study to detail the specific aspects of the actions all the proposed stakeholders should be involved with.

We discuss each of these priority actions in turn below.

59 DBT et al. (2023).

Action 1: Link up with relevant stakeholders to proactively track and analyse the landscape of tools for trustworthy AI in the UK, the US and beyond

Given the rapidly evolving capabilities of AI, the many ongoing global conversations about AI oversight, and the UK's aim to take on a strategic, international leadership role in AI,[60] we propose that in the short term, the UK adopts a pro-active role in continuously tracking and monitoring the potentially fragmented tools for trustworthy AI landscape. Given the pace at which AI is developing, it is important that the UK remains on the front foot so that it does not fall behind the developments – both technical and regulatory – that are taking place in the wider tools for trustworthy AI ecosystem.

As noted in this report, the OECD has created an online, interactive platform – the Catalogue of Tools & Metrics for Trustworthy AI – 'to share and compare tools and build upon each other's effort'.[61] The UK and the US could continuously cooperate with the OECD team responsible for maintaining the Catalogue to extract a more detailed understanding about the UK and the US ecosystems (and other relevant jurisdictions), to seek guidance and insights on the state of play and direction of travel, and to collaborate on the technical infrastructure and capabilities required to monitor trends (e.g. automating the data collection). Over time, this could lead to acquiring a more robust, evidence-based awareness and understanding of the wider global landscape of tools for trustworthy AI, and its implications for the UK AI market and UK–US alignment. In the UK, DSIT could continue to play an active role in this engagement, as we recognise that – through its Portfolio of AI Assurance Techniques[62] – it has partnered with the OECD.[63] As noted on the Portfolio of AI Assurance Techniques website, the current examples of AI assurance techniques will be regularly updated over time with additional case studies.[64] Furthermore, continuing to link up with local stakeholders in the wider ecosystem working on other aspects of tools for trustworthy AI – for example, the Alan Turing Institute, the British Standards Institution (BSI) and the National Physical Laboratory in the UK,[65] as well as universities[66] – will help cover a broader range of tools and ensure a more holistic understanding of the environment and its development trajectory.

Potential stakeholders to involve: DSIT, including the Responsible Technology Adoption Unit (RTA); techUK;[67] the AI Standards Hub; the OECD; and the US National Institute of Standards and Technology (NIST).

60 DSIT (2024b).

61 OECD (2024a).

62 The portfolio was initially developed by the Centre for Data Ethics and Innovation. On 6 February 2024, this centre changed its name to the Responsible Technology Adoption Unit (RTA): CDEI and DSIT (2024c).

63 CDEI and DSIT (2024a); OECD (2024a).

64 CDEI and DSIT (2024b).

65 These three organisations, with the support of the UK government, are involved in a joint initiative – the AI Standards Hub – with a mission to 'advance trustworthy and responsible AI with a focus on the role that standards can play as governance tools and innovation mechanisms' (AI Standards Hub 2024).

66 As noted in Chapter 2, based on the examples of tools identified, there appears to be a greater degree of collaboration between industry and academia in the US compared with the UK.

67 They are included because they were involved in the initial development of the Portfolio of AI Assurance Techniques.

Action 2: Systematically capture experiences and lessons learnt on tools for trustworthy AI, share insights with stakeholders, and use them to anticipate potential future directions

As an extension of Action 1, wider sharing of information and analysis on tools for trustworthy AI between the UK and the US – historical, current and planned – can help stakeholders avoid 'reinventing the wheel', particularly for smaller actors, such as small and medium-sized enterprises and not-for-profits, which might be resource constrained or lack relevant expertise. In the near term, this could enable the UK to get a more informed sense of what is happening 'on the ground' with respect to tool development in the UK and the US, beyond tracking developments (as outlined in Action 1).

In addition to capturing descriptive information on specific case studies of tools being collated through participative mechanisms and outreach activities with US and UK stakeholders developing and using those tools, researchers could also capture and systematically curate information on such aspects as drivers, barriers, experiences (including what works and does not work) and good practices. This information could then be disseminated in a transparent and accessible manner (e.g. through websites or through workshops and webinars) to relevant stakeholders as a primary shared resource that could include, for example, a publicly available case study bank (showcasing specific examples of tools as well as cross-analyses of the case studies) and toolkits (highlighting good practices, factsheets, practical and operational guidance, key players, sector-specific information, etc.).[68] This could be a living

68 This function could be (partially) served by the 'Introduction to AI assurance' resource, which is planned to be published by the UK government in Spring 2024 and aims to raise awareness on AI assurance techniques and help stakeholders increase their understanding of trustworthy AI systems (DSIT 2024b).

resource (e.g. like an online observatory and forum) that would need to be regularly updated to reflect new developments regarding tools for trustworthy AI. Since the information and analyses contained in this resource would be stakeholder driven and incorporate market-led 'signals', such a resource would have direct implications for the trajectory of the ecosystem of trustworthy AI in the UK and the US. Furthermore, the UK and the US could consider collaborating on actively soliciting the development of tools for trustworthy AI in the context of specific challenges, such as the UK Fairness Innovation Challenge.[69]

This approach of information exchange would not only facilitate continuous improvement and innovation in tools for trustworthy AI, to keep up with the rapid pace of AI development, but also provide evidence to anticipate potential future directions. This forward-looking approach could assist in the creation of more resilient and effective tools and strategies that could potentially cope with the uncertainty of fast-changing developments in AI. Together with Action 1, these activities could also contribute to increasing the awareness and accessibility of tools for trustworthy AI for stakeholders in the ecosystem. DSIT's Portfolio of AI Assurance Techniques[70] is a helpful foundation to build on and potentially expand out over time, along with the AI Standards Hub.[71] Depending on the availability of resources, the portfolio and associated activities could be co-developed with a US-based entity, such as NIST. As noted in Action 1, it would be valuable to draw on the experiences of those involved in the OECD Catalogue of Tools & Metrics for Trustworthy AI.[72]

Potential stakeholders to involve: DSIT, including the RTA and the UK AI Safety Institute (UK AISI); the AI Standards Hub; the Government Office for Science; the OECD; NIST; and the US AI Safety Institute (US AISI).

69 DSIT et al. (2024).
70 CDEI and DSIT (2024b).
71 AI Standards Hub (2024).
72 OECD (2024a).

Action 3: Promote the consistent use of a common vocabulary for trustworthy AI among stakeholders in the UK and the US

The emergence of numerous AI oversight frameworks across the world, including in the UK, the US and the European Union (EU), as noted in Chapter 1, highlights the need for developing a common taxonomy and for aligning terminology and vocabulary, particularly when it comes to operationalising a complex concept such as trustworthy AI. There is some inconsistency in terms of how various foundational concepts associated with trustworthy AI – fairness, transparency, accountability and safety, to name a few – are currently used by stakeholders in the UK, the US, and beyond (see Box 11).[73] In addition, while such terms as risk governance and risk management are defined and operationalised by entities such as standards development organisations, individual countries have their own approaches that have been developed in parallel to international efforts.[74]

This existence of parallel tracks could be problematic, as a key step in boosting effective cooperation between two notable jurisdictions, such as the UK and the US, is to ensure that there is clarity and that stakeholders involved have a shared understanding – a lexicon – of the different phrases and concepts, while considering their respective unique socio-technical and regulatory contexts.[75] This shared understanding, in turn, is key to achieving interoperability as well as regulatory clarity.[76] The US and the EU have already made progress towards developing a shared terminology and taxonomy for AI (currently covering 65 terms)[77] through the EU–US Trade and Technology Council (TTC) Joint Roadmap for Trustworthy AI and Risk Management.[78] The UK could consider leveraging this work[79] and/or getting involved to further boost transatlantic cooperation and harmonisation on AI, while tailoring it to the context of AI activities in the UK. Rather than starting from scratch, it will be helpful to draw on existing resources[80] that are concerned with taxonomies and terminologies for trustworthy AI.[81]

Potential stakeholders to involve: DSIT, including the RTA; the FCDO (British Embassy Washington); the BSI; ANSI; NIST; and the EC.

73 INT03; INT10.

74 INT10.

75 It is worth noting that while it is important to have clarity and consensus on what trustworthy AI is and what its key characteristics are, it is perhaps less necessary to establish consensus on *how* to achieve trustworthy AI. For example, it could be more valuable to seek to translate and map terminology to aid interoperability provided differing approaches are mutually understood.

76 INT10.

77 EC (2023b).

78 EC (2022, 2023a).

79 It could also draw on similar efforts conducted by the International Organization for Standardization (ISO), the Institute of Electrical and Electronics Engineers (IEEE) and NIST.

80 See, for example, Newman (2023); ISO (2021).

81 A potential venue for this effort – particularly from the perspective of the safety of advanced AI systems – is the planned *International Report on the Science of AI Safety,* which will be released by the UK government in Spring 2024 (DSIT 2024b).

Box 11: Four different definitions of fairness

The UK, the US, the EC and the OECD incorporate the concept of fairness into their conceptualisations of trustworthy AI. However, the definitions or interpretations of fairness used in all four contexts differ in subtle but important ways. We reproduce these definitions here:

UK: 'AI systems should not undermine the legal rights of individuals or organisations, discriminate unfairly against individuals or create unfair market outcomes. Actors involved in all stages of the AI life cycle should consider definitions of fairness that are appropriate to a system's use, outcomes and the application of relevant law.'[82]

US: 'Concerns for equality and equity by addressing issues such as harmful bias and discrimination.'[83]

EC: 'Fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.'[84]

OECD: 'AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.'[85]

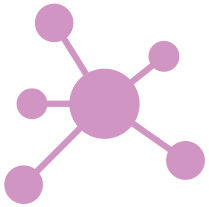
82 DSIT (2023).
83 NIST (2023).
84 EC (2019).
85 OECD (2019).

Action 4: Encourage the inclusion of assessment processes in the development and use of tools for trustworthy AI to gain a better understanding of their effectiveness

As noted in Chapter 2, there is limited evidence associated with the formal assessment and evaluation of tools for trustworthy AI in the UK and the US. For example, does each tool enhance the specific aspects of trustworthy AI it is concerned with? Across the portfolio of tools, is there a general improvement in trustworthy AI, risk management and desirable outcomes associated with the different approaches? Does increased trust persist over time? While it might be too resource intensive to formally assess the quality and effectiveness of every tool, given the sizeable number and diverse range of tools being developed and used, the UK and the US could potentially consider informally or formally assessing and cross-analysing subsets of tools across the AI value chain (through stakeholder feedback, researcher observations, etc.). In addition, developers of tools for trustworthy AI could be encouraged to include (more) information about internal pre-release assessment and to participate in relevant post-release assessment activities.

This assessment would help not only to learn lessons (see Action 2), but also to track and understand the impacts and longer-term outcomes associated with tool design, development, deployment and use. Independent assessments would also promote transparency and accountability. Conducting longer-term follow-up studies can improve the evidence base and aid in understanding the effectiveness of tools. Over time, feedback from these assessments could contribute to helping developers design more effective and innovative tools, which can improve a wider range of outcomes. This action is directly linked to Action 2 and could involve a set of follow-on activities that are rolled out over time.

Potential stakeholders to involve: DSIT, including the RTA and UK AISI; the British Embassy Washington; NIST; the Evaluation Task Force in the UK; evaluation practitioners; and US AISI.



Action 5: Continue to partner and build diverse coalitions with international organisations and initiatives, and to promote interoperable tools for trustworthy AI

While tools for trustworthy AI are important, they are not enough on their own, particularly given the rapid proliferation of AI governance–related activities across the globe. Trustworthy AI is a global and cross-sectoral issue that requires the collaboration and coordination of AI actors from different countries, regions, sectors and disciplines, involving stakeholders with diverse skills, to share best practices, learn from each other and harmonise tools for trustworthy AI while respecting each other's unique contexts. To ensure the development of responsible and trustworthy AI – and consequently of tools for trustworthy AI – the UK and the US could continue to engage with various stakeholders and promote inclusive dialogue and information exchange that involves diverse perspectives, with a particular emphasis on multistakeholder initiatives, international organisations, and other countries and regions that share similar values and a similar vision.

Examples of organisations and initiatives include: the Global Partnership on Artificial Intelligence (GPAI)[86]; OECD.AI[87]; the UN's High-Level Advisory Body on Artificial Intelligence[88]; UNESCO[89]; the Hiroshima AI Process[90] and other AI-related G7 activities; various AI-related activities across the EU[91] (including, for example, the EC's proposed legal framework on AI – the 'AI Act')[92]; and outcomes of the AI Safety Summit 2023[93] (e.g. partnerships with AI Safety Institutes across the globe).[94] The UK and the US are already actively involved to varying degrees in these and other multilateral fora. It may also be useful to draw on the lessons learnt from developing recent transatlantic strategic collaborative vehicles in other related contexts, such as biosecurity[95] and cybersecurity.[96] Against the backdrop of the current regulatory uncertainty around AI, these high-level collaborative set-ups will foster dialogue and cooperation on the global governance and coordination of AI, as well as provide avenues for the adoption and implementation of the principles and practices of trustworthy AI – and, subsequently, the development and deployment of tools for trustworthy AI – in a compatible and interoperable manner.

Potential stakeholders to involve: DSIT, including the RTA and UK AISI; the FCDO; NIST; US AISI; the OECD; the EC; and the UN.

86 GPAI (2024).
87 OECD (2024c).
88 United Nations, Office of the Secretary-General's Envoy on Technology (2024).
89 UNESCO (2024).
90 The White House (2023b).
91 EC (2024b).
92 EC (2024a).
93 FCDO et al. (2023).
94 These high-level international collaborations can be further developed through partnerships between dedicated institutes in the US, the UK and other countries. For example, the UK AISI has formed a partnership with the US AISI and the Singaporean government to collaborate on safety testing of AI models (DSIT 2024b). As a further signal of strong bilateral collaboration between the UK and US on AI safety, on 1 April 2024, a memorandum of understanding was signed to enable the UK and US AISIs 'to work closely to develop an interoperable programme of work and approach to safety research, to achieve their shared objectives on AI safety' (UK AISI 2024).
95 Cabinet Office (2024).
96 NCSC (2023).

Action 6: Join forces to provide resources such as data and computing power to support and democratise the development of tools for trustworthy AI

The current generation of foundation models are large and complex and are trained on vast amounts of publicly available data. This means it will become harder to find data that can be used as a holdout data set.[97] This store of non-synthetic data that is not included in any existing models could be a useful resource for the development of tools to measure trustworthiness.[98] There is potentially much unique data within governments that is not being accessed.[99] The recently created US National AI Research Resource (NAIRR) and the UK AI Research Resources (AIRR) could potentially help provide access to these data through a joint cloud service. Similarly, developing and deploying large foundation models and creating appropriate tools for ensuring the trustworthiness of these models can require large amounts of compute. Academic and not-for-profit research can be an important source of independent research on AI assurance techniques. However, these researchers are being 'priced out' of this research because of the steep cost of compute.[100] NAIRR and AIRR, together with the US national labs, could potentially help provide the necessary compute capacity for these efforts.[101] Furthermore, it may be helpful to draw on the experiences of current models of international collaboration in AI compute, such as the recently announced memorandum of understanding between the UK and Canada.[102] Directing efforts towards more equitable access and democratising compute and data to 'internationalise' tools for trustworthy AI could not only address the UK–US landscape, but also point towards common ambitions across key multilateral fora in the wider AI governance ecosystem (as highlighted in Action 5).

Potential stakeholders to involve: NAIRR; AIRR; DSIT, including UK AISI; UKRI; the US National Science Foundation (NSF); and US national labs.

In Figure 2, we provide a visual summary of the six practical actions suggested for policymakers.

97 In machine learning, a holdout dataset refers to data that has never been used in the training of the model. These types of data can be used to independently validate certain characteristics of the model while avoiding the need to use data that the model is familiar with. Since very large foundation models are trained on almost all publicly available data, holdout data can become increasingly hard to find: Raschka (2018).

98 Synthetic data is data that has been generated through an algorithm; non-synthetic data is data that has been measured and collected in the 'real world': Jordon et al. (2022).

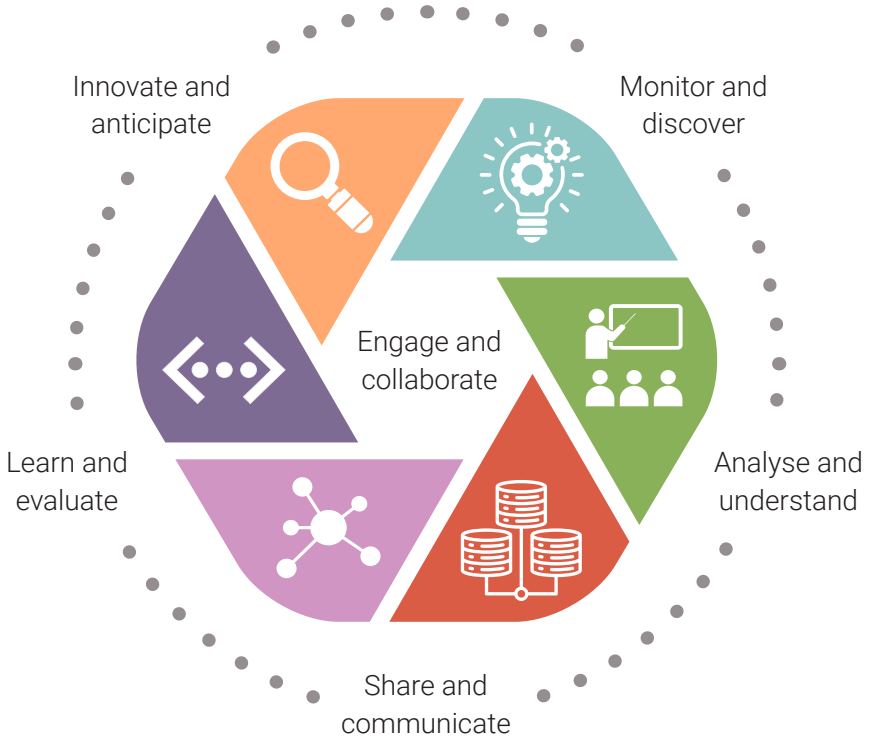
99 INT06. An example of a move to address this is the collaboration between NASA and IBM to release NASA's Harmonized Landsat and Sentinel-2 (HLS) dataset of geospatial data: Blumenfeld (2023).

100 INT05; INT09.

101 UKRI (2024); NSF (2024).

102 DSIT (2024a).

Figure 2. Practical considerations for UK and US policymakers to help build a linked-up, aligned and agile ecosystem



ACTION 1

Link up with relevant stakeholders to proactively track and analyse the landscape of tools for trustworthy AI in the UK, the US and beyond

ACTION 2

Systematically capture experiences and lessons learnt on tools for trustworthy AI, share those insights with stakeholders and use them to anticipate potential future directions

ACTION 3

Promote the consistent use of a common vocabulary for trustworthy AI among stakeholders in the UK and the US

ACTION 4

Encourage the inclusion of assessment processes in the development and use of tools for trustworthy AI to gain a better understanding of their effectiveness

ACTION 5

Continue to partner and build diverse coalitions with international organisations and initiatives, and to promote interoperable tools for trustworthy AI

ACTION 6

Join forces to provide resources such as data and computing power to support and democratise the development of tools for trustworthy AI

Potential stakeholders to involve across the different actions: Department for Science, Innovation and Technology (including the Responsible Technology Adoption Unit and UK AI Safety Institute); Foreign, Commonwealth & Development Office (including the British Embassy Washington); AI Standards Hub; UK Research and Innovation; AI Research Resource; techUK; Evaluation Task Force in the UK; Government Office for Science; National Institute of Standards and Technology; US AI Safety Institute; National Science Foundation; National Artificial Intelligence Research Resource; US national laboratories; Organisation for Economic Co-operation and Development; European Commission; United Nations (and associated agencies); standards development organisations.

Source: RAND Europe analysis

Bibliography

AI Standards Hub (homepage). 2024. As of 5 February 2024: https://aistandardshub.org/the-ai-standards-hub/

Berditchevskaia, Aleks, Eirini Malliaraki, & Kathy Peach. 2021. *Participatory AI for Humanitarian Innovation*. London: NESTA. As of 5 February 2024: https://media.nesta.org.uk/documents/Nesta_Participatory_AI_for_humanitarian_innovation_Final.pdf

Blumenfeld, Josh. 2023. 'NASA and IBM Openly Release Geospatial AI Foundation Model for NASA Earth Observation Data.' As of 6 February 2024: https://www.earthdata.nasa.gov/news/impact-ibm-hls-foundation-model

Cabinet Office. 2024. 'UK and U.S. Announce New Strategic Partnership to Tackle Increased Biological Threats.' London: HM Government. As of 5 February 2024: https://www.gov.uk/government/news/uk-and-us-announce-new-strategic-partnership-to-tackle-increased-biological-threats

CDEI (Centre for Data Ethics and Innovation) and DSIT (Department for Science, Innovation, and Technology). 2024a. 'Find Out About Artificial Intelligence (AI) Assurance Techniques.' As of 22 January 2024: https://www.gov.uk/ai-assurance-techniques

———. 2024b. 'Portfolio of AI Assurance techniques.' As of 22 January 2024: https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques

———. 2024c. 'The CDEI is Now the Responsible Technology Adoption Unit.' As of 6 February 2024: https://www.gov.uk/government/news/the-cdei-is-now-the-responsible-technology-adoption-unit

Comcast. 2023. 'Project Guard Rail.' As of 22 January 2024: https://github.com/Comcast/ProjectGuardRail

Cortes, Corinna, Xavier Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, & Scott Yang. 2017. 'AdaNet: Adaptive Structural Learning of Artificial Neural Networks.' As of 5 February 2024: http://proceedings.mlr.press/v70/cortes17a.html

Covert, Ian, Scott Lundberg, & Su-In Lee. 2020. 'Understanding Global Feature Contributions with Additive Importance Measures.' As of 5 February 2024: https://arxiv.org/abs/2004.00668

DBT (Department for Business & Trade), FCDO (Foreign, Commonwealth and Development Office), & Prime Minister's Office. 2023. 'The Atlantic Declaration.' London: HM Government. As of 22 January 2024: https://www.gov.uk/government/publications/the-atlantic-declaration/the-atlantic-declaration

DSIT (Department for Science, Innovation and Technology). 2023. 'A Pro-innovation Approach to AI Regulation.' London: HM Government. As of 22 January 2024: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

———. 2024a. 'UK-Canada Cooperation in AI Compute: Memorandum of Understanding.' London: HM Government. As of 5 February 2024: https://www.gov.uk/government/publications/uk-canada-cooperation-in-ai-compute-memorandum-of-understanding/uk-canada-cooperation-in-ai-compute-memorandum-of-understanding

———. 2024b. 'A Pro-innovation Approach to AI Regulation: Government Response' London: HM Government. As of 6 February 2024: https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response

DSIT (Department for Science, Innovation and Technology), Innovate UK, the Equality and Human Rights Commission and the Information Commissioner's Office. 2024. 'Fairness Innovation Challenge.' As of 5 February 2024: https://fairnessinnovationchallenge.co.uk/

EC (European Commission). 2019. 'Ethics Guidelines for Trustworthy AI.' As of 22 January 2024: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

———. 2022. 'TTC Joint Roadmap for Trustworthy AI and Risk Management.' As of 22 January 2024: https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management

———. 2023a. 'EU–US Trade and Technology Council.' As of 22 January 2024: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/stronger-europe-world/eu-us-trade-and-technology-council_en

———. 2023b. 'EU-U.S. Terminology and Taxonomy for Artificial Intelligence.' As of 22 January 2024: https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence

———. 2024a. 'AI Act.' As of 22 January 2024: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

———. 2024b. 'European Approach to Artificial Intelligence.' As of 22 January 2024: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

FCDO (Foreign, Commonwealth and Development Office), DSIT (Department for Science, Innovation and Technology), & UK AISI (UK AI Safety Institute). 2023. 'AI Safety Summit 2023.' As of 11 April 2024: https://www.gov.uk/government/topical-events/ai-safety-summit-2023

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christop Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossie, Burkhard Schafer, Peggy Valcke, & Effy Vayena. 2019. *AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.* Brussels: AI4People. As of 11 April 2024: https://www.eismd.eu/wp-content/uploads/2019/11/AI4People's-Ethical-Framework-for-a-Good-AI-Society_compressed.pdf

GPAI (Global Partnership on Artificial Intelligence). 2024. 'Global Partnership on Artificial Intelligence.' As of 22 January 2024: https://gpai.ai/

Graham, Logan, Abigail Gilbert, Joshua Simons, Anna Thomas, & Helen Mountfield. 2020. *Artificial Intelligence in Hiring: Assessing Impact on Equality.* London: Institute for the Future of Work. As of 5 February 2024: https://www.ifow.org/publications/artificial-intelligence-in-hiring-assessing-impacts-on-equality

Gupta, Abhishek, Camylle Lanteigne, & Sara Kingsley. 2020. 'SECure: A Social and Environmental Certificate for AI Systems.' As of 5 February 2024: https://arxiv.org/abs/2006.06217

IBM Research. 2024a. 'AI Fairness 360.' As of 22 January 2024: https://aif360.res.ibm.com/

———. 2024b. 'AI Privacy 360.' As of 22 January 2024: https://aip360.res.ibm.com/

———. 2024c. 'AI Explainability 360.' As of 22 January 2024: https://aix360.res.ibm.com/

———. 2024d. 'Uncertainty Quantification 360.' As of 22 January 2024: https://uq360.res.ibm.com/

———. 2024e. 'AI FactSheets 360.' As of 22 January 2024: https://aifs360.res.ibm.com/

———. 2024f. 'Adversarial Robustness 360 – Resources.' As of 22 January 2024: https://art360.res.ibm.com/resources#overview

ISO (International Organization for Standardization). 2021. *ISO/IEC DIS 22989(en) Information technology — Artificial intelligence — Artificial Intelligence Concepts and Terminology.* Geneva: International Organization for Standardisation. As of 5 February 2024: https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:dis:ed-1:v1:en

Jordon, James, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel Cohen, & Adrian Weller. 2022. *Synthetic Data – What, Why and How?* London: Alan Turing Institute. As of 5 February 2024: https://arxiv.org/abs/2205.03257

Kearns, Michael, Seth Neel, Aaron Roth, & Zhiwei Steven Wu. 2018. 'Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness.' As of 5 February 2024: https://arxiv.org/abs/1711.05144

Kumar, Ram Shankar Siva. 2021. 'AI Security Risk Assessment Using Counterfit.' Redmond, WA: Microsoft. As of 5 February 2024: https://www.microsoft.com/en-us/security/blog/2021/05/03/ai-security-risk-assessment-using-counterfit/

NCSC (National Cyber Security Centre). 2023. 'Guidelines for Secure AI System Development.' As of 5 February 2024: https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

Newman, Jessica. 2023. *A Taxonomy of Trustworthiness for Artificial Intelligence.* Berkeley, CA: University of California Berkeley Centre for Long-Term Cybersecurity. As of 5 February 2024: https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf

NIST (National Institute of Standards and Technology). 2023. *Artificial Intelligence Risk Management Framework (AIRMF1.0).* Gaithersburg, MD: National Institute of Standards and Technology. As of 22 January 2024: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

NSF (National Science Foundation). 2024. 'Democratizing the Future of AI R&D: NSF to Launch National AI Research Resource Pilot.' As of 5 February 2024: https://new.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai

OECD (Organisation for Economic Co-operation and Development). 2019. 'Recommendation of the Council on Artificial Intelligence.' As of 22 January 2024: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

———. 2021. 'Tools for Trustworthy AI: A Framework to Compare Implementation Tools for Trustworthy AI.' As of 22 January 2024: https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm

———. 2024a. 'Catalogue of Tools & Metrics for Trustworthy AI – About the Catalogue.' As of 22 January 2024: https://oecd.ai/en/catalogue/faq

———. 2024b. ‘Catalogue of Tools & Metrics for Trustworthy AI – Show Use Cases.’ As of 22 January 2024: https://oecd.ai/en/catalogue/tool-use-cases

———. 2024c. ‘Policies, Data and Analysis for Trustworthy Artificial Intelligence.’ As of 19 January 2024: https://oecd.ai/en/

———. 2024d. ‘Catalogue of Tools & Metrics for Trustworthy AI – Show Tools.’ As of 22 January 2024: https://oecd.ai/en/catalogue/tools

PAI (Partnership on AI). 2023. ‘PAI’s Guidance for Safe Foundation Model Deployment.’ As of 22 January 2024: https://partnershiponai.org/modeldeployment/

Raschka, Sebastian. 2018. ‘Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.’ As of 5 February 2024: https://arxiv.org/abs/1811.12808

Rolls Royce. 2023. ‘The Aletheia Framework.’ As of 5 February 2024: https://www.rolls-royce.com/innovation/the-aletheia-framework.aspx

The White House. 2022. ‘Blueprint for an AI Bill of Rights.’ As of 22 January 2024: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

———. 2023a. ‘Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.’ As of 11 April 2024: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

———. 2023b. ‘G7 Leaders’ Statement on the Hiroshima AI Process.’ As of 19 January 2024: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/g7-leaders-statement-on-the-hiroshima-ai-process/

UK AISI (UK AI Safety Institute). 2024. ‘Collaboration on the safety of AI: UK-US memorandum of understanding.’ London: HM Government. As of 11 April 2024: https://www.gov.uk/government/publications/collaboration-on-the-safety-of-ai-uk-us-memorandum-of-understanding/collaboration-on-the-safety-of-ai-uk-us-memorandum-of-understanding

UKRI (UK Research and Innovation). 2024. ‘AI Research Resource Funding Opportunity Launches.’ As of 5 February 2024: https://www.ukri.org/news/ai-research-resource-funding-opportunity-launches/

UNESCO. 2021. ‘Recommendation on the Ethics of Artificial Intelligence.’ As of 22 January 2024: https://unesdoc.unesco.org/ark:/48223/pf0000380455

———. 2024. ‘Artificial Intelligence.’ As of 19 January 2024: https://www.unesco.org/en/artificial-intelligence

United Nations, Office of the Secretary-General’s Envoy on Technology. 2024. ‘High-Level Advisory Body on Artificial Intelligence.’ As of 19 January 2024: https://www.un.org/techenvoy/ai-advisory-body

Annex A. Further details on the underlying principles of trustworthy AI from different stakeholders

In this annex, we provide further details on the underlying principles and characteristics of trustworthy AI from different stakeholders. In the sections below, we reproduce the interpretations of trustworthy AI according to four sources.

A.1. UK government

The UK government approach to trustworthy AI is based on the following principles[103]:

- **Safety, security and robustness**: 'AI systems should function in a robust, secure and safe way throughout the AI life cycle, and risks should be continually identified, assessed and managed.'
- **Appropriate transparency and explainability:** 'Transparency refers to the communication of appropriate information about an AI system to relevant people (for example, information on how, when, and for which purposes an AI system is being used). Explainability refers to the extent to which it is possible for relevant parties to access, interpret and understand the decision-making processes of an AI system.'
- **Fairness:** 'AI systems should not undermine the legal rights of individuals or organisations, discriminate unfairly against individuals or create unfair market outcomes. Actors involved in all stages of the AI life cycle should consider definitions of fairness that are appropriate to a system's use, outcomes and the application of relevant law.'
- **Accountability and governance:** 'Governance measures should be in place to ensure effective oversight of the supply and use of AI systems, with clear lines of accountability established across the AI life cycle.'
- **Contestability and redress:** 'Where appropriate, users, impacted third parties and actors in the AI life cycle should be able to contest an AI decision or outcome that is harmful or creates material risk of harm.'

A.2 National Institute of Standards and Technology (US)

For AI systems to be trustworthy, they often need to be responsive to a multiplicity of criteria that are of value to interested parties. Approaches which enhance AI trustworthiness can reduce negative AI risks. The NIST Artificial Intelligence Risk Management Framework articulates the following characteristics of trustworthy AI and offers guidance for addressing them[104]:

- **Validity:** 'Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.'

103 DSIT (2023).
104 NIST (2023).

- **Reliability:** 'Ability of an item to perform as required, without failure, for a given time interval, under given conditions.'
- **Safety:** 'AI systems should not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered.'
- **Resilience:** The ability to 'withstand unexpected adverse events or unexpected changes in the environment or use' of AI systems – or the ability to 'maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary.'
- **Transparency:** 'The extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so.'
- **Explainability:** 'A representation of the mechanisms underlying AI systems' operation.'
- **Interpretability:** 'The meaning of AI systems' output in the context of their designed functional purposes.'
- **Privacy:** 'Refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity.'
- **Fairness:** 'Concerns for equality and equity by addressing issues such as harmful bias and discrimination.'

'Creating trustworthy AI requires balancing each of these characteristics based on the AI system's context of use. While all characteristics are socio-technical system attributes, accountability and transparency also relate to the processes and activities internal to an AI system and its external setting. Neglecting these characteristics can increase the probability and magnitude of negative consequences.'[105]

A.3. European Commission

The EC asked a High-Level Expert Group on AI to provide advice on the EU's strategy for AI. One of the tasks of the group was to draft Ethics Guidelines for Trustworthy Artificial Intelligence. According to the expert group, AI should be[106]:

- **Lawful:** 'Complying with all applicable laws and regulations.'
- **Ethical:** 'Ensuring adherence to ethical principles and values.'
- **Robust:** 'AI systems should offer a consistent performance regardless of the context or data.'

There are four ethical principles, rooted in fundamental rights, that are 'ethical imperatives' that must be respected at all times[107]:

- **Respect for human autonomy:** 'Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills.'

105 NIST (2023).
106 EC (2019).
107 EC (2019).

- **Prevention of harm:** 'AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure.'
- **Fairness:** 'Fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.… The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.'
- **Explicability:** 'Processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected.'

In order to meet these principles, AI systems should at least meet these seven requirements:

- **Human agency and oversight:** 'AI systems should support human autonomy and decision making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights and allow for human oversight.'
- **Technical robustness and safety:** 'Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm and preventing unacceptable harm.'
- **Privacy and data governance:** 'Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.'
- **Transparency:** 'The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. The processes and decisions made by AI should be explainable. AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system.'
- **Diversity, non-discrimination and fairness:** 'In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment.'
- **Environmental and societal well-being:** 'In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as for instance the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations.'

- **Accountability:** 'The requirement of accountability complements the above requirements and is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.'

A.4. OECD

The OECD AI principles were adopted by the OECD Council on Artificial Intelligence in 2019, with the goal of promoting the responsible stewardship of AI. They include the following value-based principles[108]:

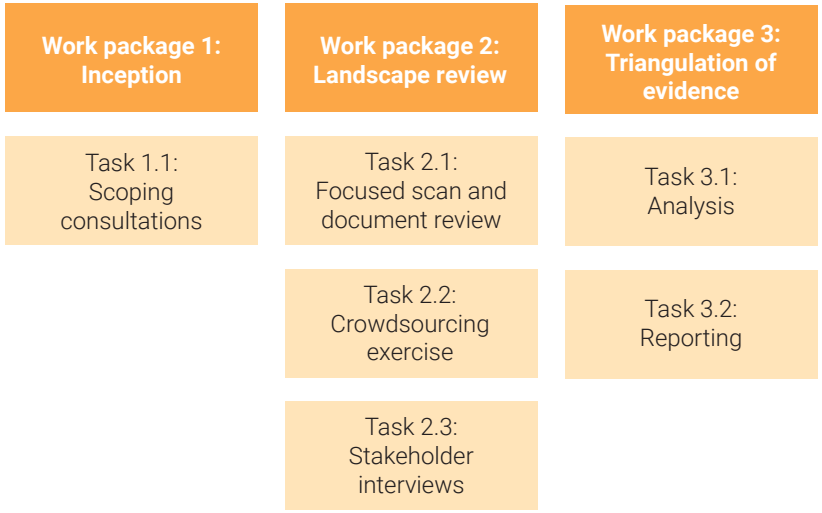
- **Inclusive growth, sustainable development and well-being:** 'Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.'
- **Human-centred values and fairness:** 'AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.'
- **Transparency and explainability:** 'AI actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art to foster a general understanding of AI systems, to make stakeholders aware of their interactions with AI systems, including in the workplace, to enable those affected by an AI system to understand the outcome, and to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.'
- **Robustness, security and safety:** 'AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.'
- **Accountability:** 'AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.'

108 OECD (2019).

Annex B. **Detailed methodological approach**

In this section, we provide a detailed summary of the study's methodology, along with the key caveats of our analysis. As shown in the figure below, the research was split into three work packages. We describe each of the work packages in the sections below.

Figure 3. Overall research approach and associated methodologies



B.1. Work Package 1 (WP1): Inception

B.1.1. Task 1.1: Scoping consultations

After an inception meeting with the British Embassy Washington team, we conducted three targeted scoping consultations – by email or Microsoft Teams – with experts in the RAND Corporation with knowledge of tools or initiatives for trustworthy AI. These consultations aimed to establish a baseline understanding of, and insights specific to, the tools for a trustworthy AI landscape, focusing on developments in the UK and the US. The scoping consultations were also used to obtain suggestions for stakeholder interviewees and articles to consult in the next work package.

B.2. Work Package 2 (WP2): Landscape review

B.2.1. Task 2.1: Focused scan and document review

Building on the insights from the scoping consultations, we carried out a focused review of the literature and data sources to collate and synthesise information about tools and initiatives for trustworthy AI that are being developed, deployed and used in the UK and the US. The review aimed to better understand the range of tools and initiatives that have been developed or implemented, how the tools work, the type of tool, the AI dimensions and characteristics they cover, and the target audience. To identify a longlist of tools that characterise the respective ecosystems in the UK and the US, we consulted different databases – notably, these included the OECD's Catalogue of Tools & Metrics for Trustworthy

AI,[109] DSIT's Portfolio of AI Assurance Techniques[110] and a selection of toolboxes and toolkits developed by technology companies.[111] In addition, we conducted a series of targeted searches in Google. The review collected a wide-ranging set of relevant information on different actors in the UK and the US AI ecosystems, including research organisations and universities, thinktanks, national and regional industry associations, and government and industry initiatives on AI. Specifically, we extracted the following information about each tool into an Excel spreadsheet (the final longlist of tools we compiled is presented in the accompanying Excel file):

- Name of tool
- Short description of tool
- Developer(s) of tool
- Country or countries of tool
- Time period of development
- Type of tool
- Aim of tool
- Development stage of tool
- Target audience.

B.2.2. Task 2.2: Crowdsourcing exercise

We carried out a targeted online crowdsourcing exercise with experts to collect additional examples of tools for trustworthy AI in the UK and the US that might not have been picked up Task 2.1. We also used the exercise to ask respondents for suggestions about other useful sources of information to consult (e.g. organisations, reports, articles), as well their views on efforts towards improving collaboration between the UK and the US on trustworthy AI. The online crowdsourcing exercise was set up to run in the background once the research began, and it ran for entire duration of the study. We created a data collection template using Google Sheets that contained the main fields we wanted to capture from the experts. The exercise was primarily aimed at AI researchers and representatives from government, industry and third sector organisations. We drew on the expertise within RAND and our wider networks to compile a list of 64 stakeholders from the US, the UK and EU countries, who were invited to fill out the crowdsourcing template. In total, we received ten responses to the crowdsourcing exercise.

B.2.3. Task 2.3: Stakeholder interviews

We conducted interviews with a range of stakeholders involved in the tools for trustworthy AI ecosystem.[112] These included stakeholders connected to some of the tools we had identified in Task 2.1, as well as more general experts with knowledge of the wider landscape of tools for trustworthy AI. We conducted ten semi-structured interviews in total, covering both US and UK stakeholders from academia, industry, government and the third sector. The interviews lasted between 30 and 60 minutes and were conducted online, through Microsoft Teams. We developed a concise, tailored interview protocol that built on emerging findings from the desk research in Task 2.1. Where appropriate, we

109 OECD (2024d).

110 CDEI and DSIT (2024a).

111 These included, for example, the Microsoft Responsible AI Toolbox (https://responsibleaitoolbox.ai/), the various IBM Research AI toolkits (https://research.ibm.com/topics/trustworthy-ai), and the Google Explainable AI toolkit (https://cloud.google.com/explainable-ai).

112 As noted previously, the evidence from the interviews has been anonymised and cited throughout the report using unique interviewee identifiers (INT01, INT02, etc.).

modified the questions we asked based on the interviewee's expertise and background. Below we list the indicative topics we discussed with interviewees:

- Understanding of the phrase 'trustworthy AI'.
- Information about specific tools for making AI trustworthy.
- Awareness of wider developments and trends in the trustworthy AI space taking place in the UK and the US.
- Views on challenges associated with developing, deploying and using tools for trustworthy AI.
- Awareness of gaps or challenges in the current cooperation between the UK and the US on tools for trustworthy AI.
- Ideas for initiatives or wider priority areas to consider for future UK–US collaboration on trustworthy AI.
- Suggestions for organisations to speak to or of further resources to consult in the research.

B.3. Work Package 3 (WP3): Triangulation of evidence

B.3.1. Task 3.1: Analysis

We compiled a longlist of tools for trustworthy AI into a comprehensive database in Excel (see the accompanying Excel file) based on research conducted in all the preceding tasks (i.e. document and data review, expert crowdsourcing, stakeholder interviews). We cleaned and harmonised the information in the database and filled in, where possible, any gaps in information. We then cross-analysed the data to pull out common themes and trends associated with the tools and, where relevant, notable divergences as well. Alongside this, we analysed the interview data and integrated relevant insights and information from the interviews into the cross-analysis of the tools database. The cross-analysis of the evidence was conducted through discussions among core members of the study team. Informed by the analysis of the evidence, we also articulated a series of considerations for policymakers involved in the trustworthy AI ecosystem in the UK and the US.

B.3.2. Task 3.2: Reporting

In the final stage of the research, we synthesised all the data from the preceding stages of research. This information formed the basis of the findings included in this report. We have used message-led headings in the main sections of this report (Chapters 2 and 3) to communicate the findings of the research in a succinct manner that may be suitable for non-expert readers. Where relevant, we have also included examples of tools for trustworthy AI in the UK and the US to illustrate specific findings. In Annex C, we include the longlist of tools identified to enable readers to look up information about specific examples of tools in more depth.

B.4. Limitations of the analysis

The analysis presented in this report is subject to some caveats related to the research approach, the scope of the evidence consulted, and the analysis we undertook. These are summarised below and should be considered while interpreting the findings presented in this report.

First, the study had to be completed within approximately eight weeks over the end-of-year holiday period in 2023–2024. We therefore had to conduct a rapid analysis of the tools for trustworthy AI landscapes in the UK and the US. Nevertheless, we ensured that we drew on comprehensive and current databases of tools that covered both geographies. We also complemented these databases with some

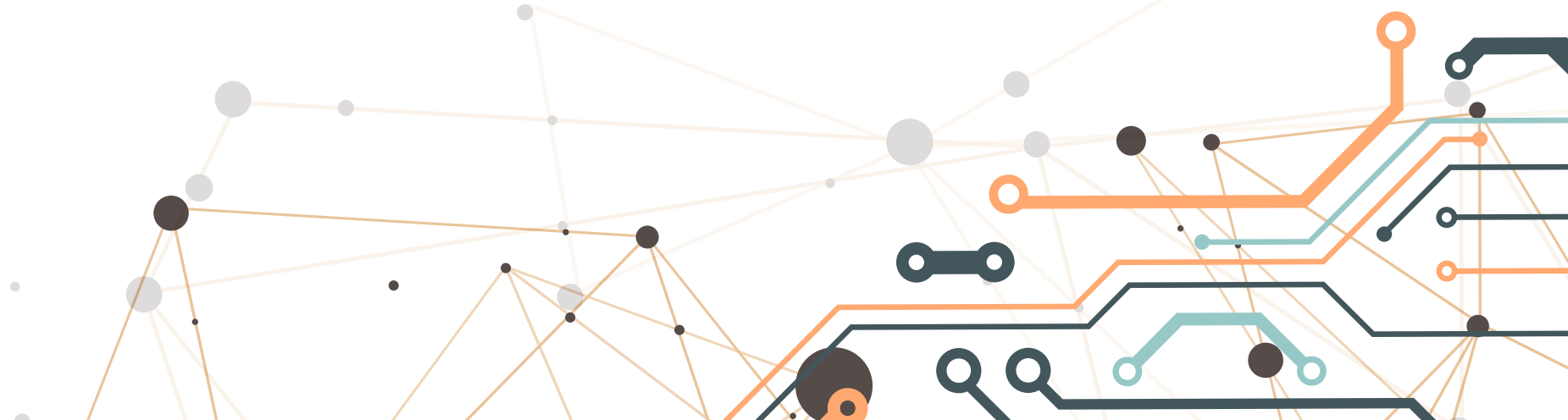
targeted searches of tools for trustworthy AI and information provided by stakeholders we interviewed.

Second, the development of trustworthy AI and linked AI governance issues is a fast-moving field, involving multiple stakeholders across the world with differing priorities. By focusing on the UK and the US, we have not included important developments taking place in this rapidly evolving field in other parts of the world (including other regulatory policy discussions). However, we are confident, based on the approach we adopted, that our analysis provides a fair and relatively holistic picture of the state of evidence (at the time of writing) in the UK and the US.

Third, while we aimed to capture as many relevant examples of tools as possible within the study timeframe, the final longlist of tools was not intended to be exhaustive or definitive, nor did we evaluate or assess the effectiveness of the tools. Rather, the examples we captured served as concrete, illustrative cases of tools that have been developed and deployed in practice to make AI trustworthy. The database of tools was intended to provide a wide-ranging snapshot of the state of play at the time of writing. Furthermore, we collated information about each tool into our database based on the information contained in the source data we consulted or using our best understanding of the information associated with the tool that we analysed. The final longlist we compiled highlights a wide spectrum of tools in this growing area that span different parts of the AI value chain, target diverse sectors, and cover a variety of dimensions and characteristics of AI trustworthiness.

Finally, we spoke to a relatively small sample of interviewees, mainly because of the tight timeframes within which the research had to be completed, which has meant that the diversity of views captured in the research is limited. Moreover, it was beyond the scope of this study to independently verify all the information that the interviewees provided. However, the interviews were only intended to complement the document and data review and to gather views and perceptions from UK- and US-based stakeholders working in the wider trustworthy AI ecosystem. Furthermore, within the sample of interviewees, we attempted to seek expert opinion across a range of stakeholders from industry, academia, government and the third sector.

Notwithstanding the caveats discussed above, we hope that the analyses and findings presented in this report will be useful to inform future thinking related to the growing and increasingly important area of tools for trustworthy AI.

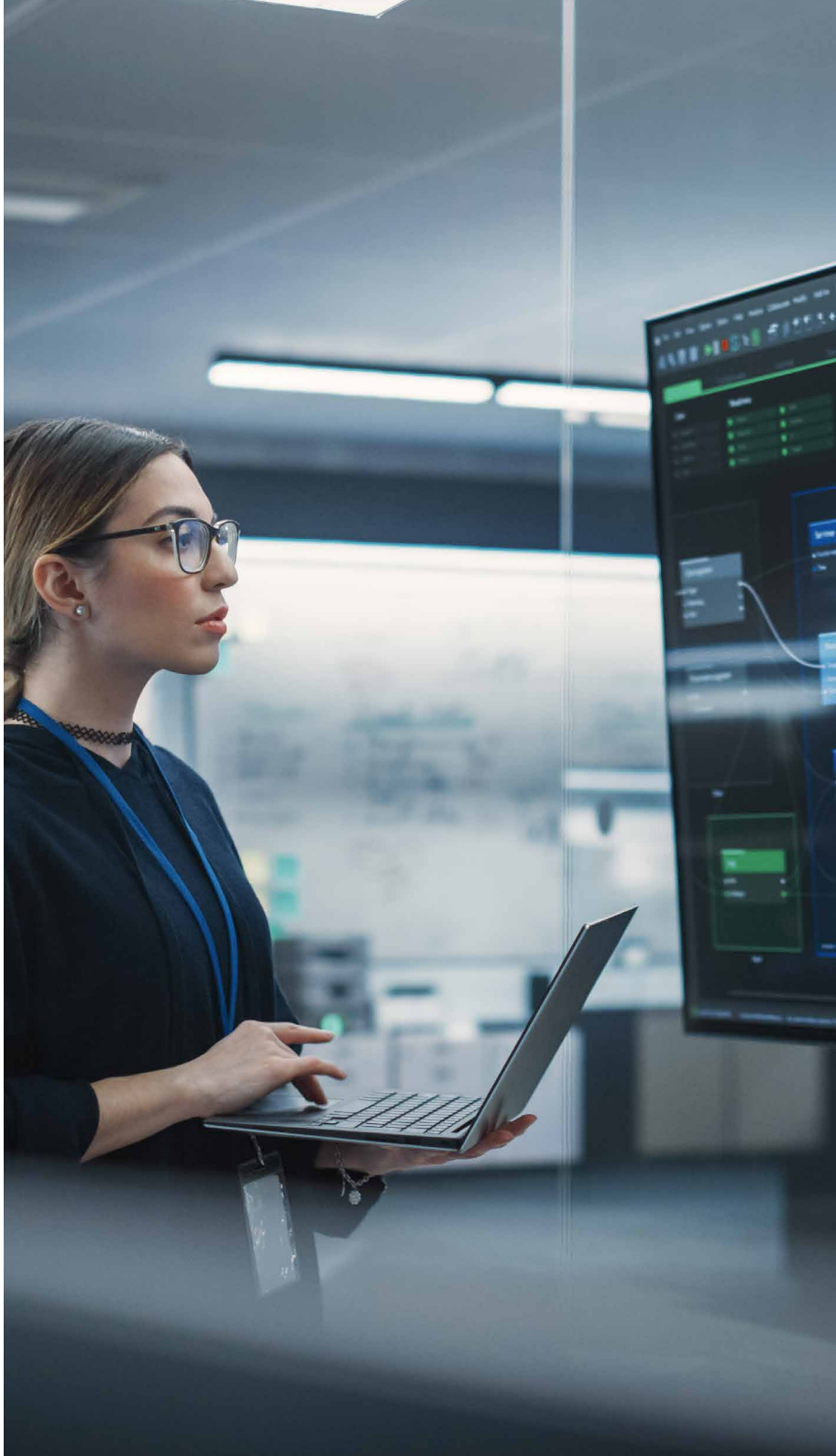


Annex C. **Longlist of tools for trustworthy AI**

This annex provides the longlist of tools identified in the research, which is presented in the accompanying Excel file. A core element of the study involved the creation of a longlist of tools for trustworthy AI in the UK and the US. The longlist was generated by collating information from existing databases of tools, targeted online searches and a crowdsourcing exercise with experts. Information associated with each tool was extracted into an Excel spreadsheet, which is attached as an accompanying Excel file. The spreadsheet is structured around the following high-level categories, each of which consists of associated sub-fields of information[113]:

- **General:** Contains data on the name and description of the tool; the developer(s); the country or countries of origin; the dates when the tool was developed and uploaded; and relevant links with further information about the tool.
- **Application:** Contains data on the tool type; its objective; the type of approach used; and its maturity. This categorisation was based on the OECD Catalogue of Tools for Trustworthy AI.[114]
- **Users:** Contains data on the target sector; the target users; and impacted stakeholders.

The longlist of tools that we had compiled at the time of writing the report (January 2024) is published as an accompanying Excel file. The Excel file should be read in conjunction with this report.



113 These fields were completed with varying levels of specificity that depended on the information associated with each tool in the underlying evidence we reviewed.

114 OECD (2024a).