EXECUTIVE SUMMARY

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.

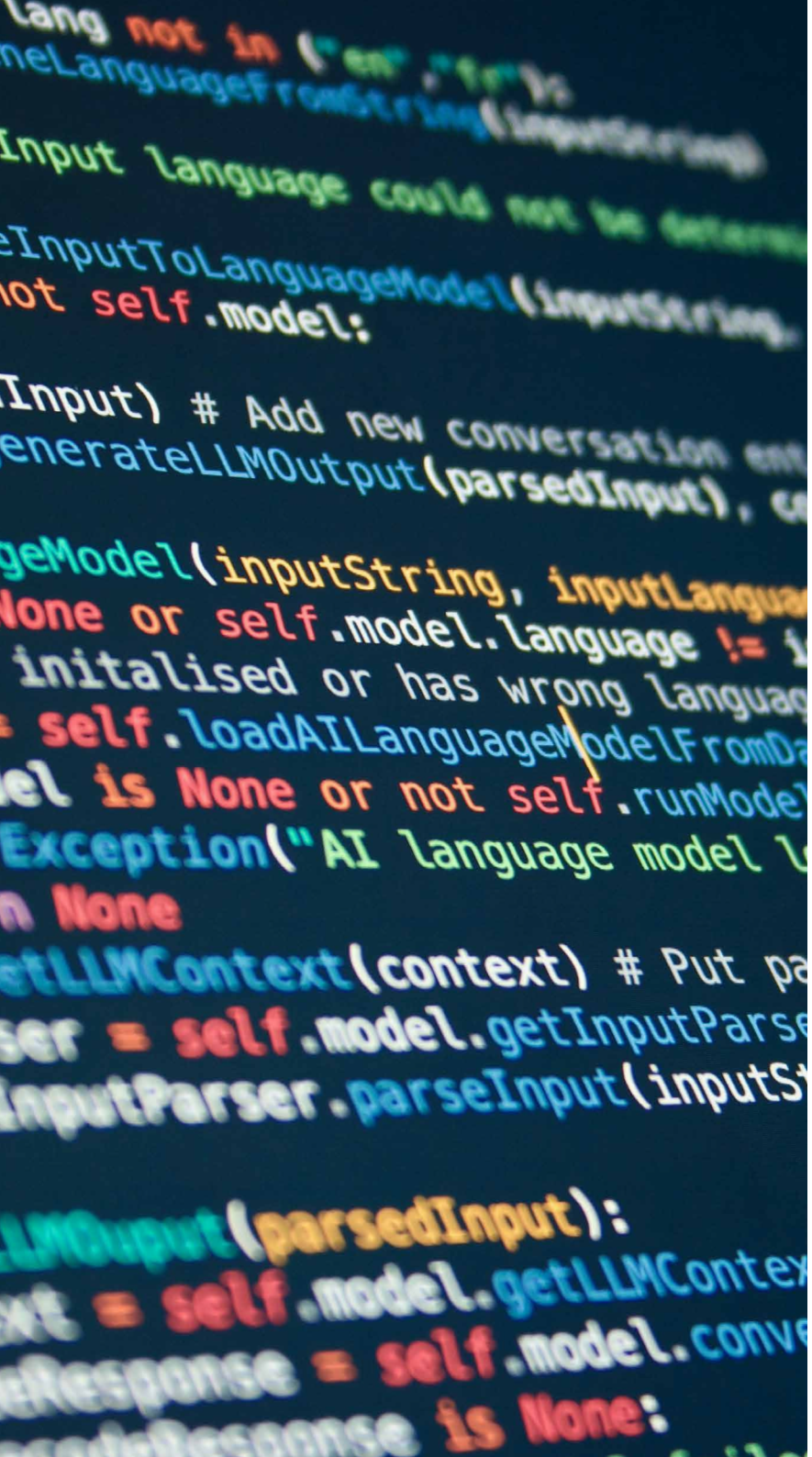
Large US technology companies are developing wide-ranging toolkits to make AI products and services more trustworthy.

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