Frameworks and Toolkits for Assuring RESPONSIBLE AI

White Paper 02, October 2025

www.rai.ac.uk

RAi UK White Paper Series

Our Mission: Translating Ideas into Impact

The Responsible Ai UK (RAi UK) White Paper Series presents interdisciplinary, thematic studies exploring how to responsibly harness the opportunities of artificial intelligence across key priority areas. Each paper aims to translate research into tangible impact.

As the national convenor of the UK's academic AI ecosystem, RAi UK brings together leading voices from **academia, government, industry**, and the **third sector** to deliver holistic assessments of the most pressing opportunities and challenges in responsible AI — and to catalyse action.

This series is designed to drive momentum by:

Convening the ecosystem, challenges, and opportunities

Collaborating with the people and organisations best placed to act

Catalysing real-world change by informing and inspiring action

Current and Forthcoming Papers

- *Responsible AI to Enable Flourishing in and by Low- and Middle-Income Countries (LMICs)*
- *Frameworks and Toolkits for Assuring Responsible AI*
- *Advancing Trustworthy Artificial Intelligence: Lessons Learned and Emerging Challenges (November 2025)*
- *Responsible AI & Healthcare* (December 2025)
- *Responsible AI & Education* (December 2025)

Have an idea for a future paper or interested in joining a future workshop?
We welcome suggestions. Get in touch: info@rai.ac.uk

Table of Contents

Executive Summary

This report explores experiences of using tools for assuring responsible practice in the build, deployment, and governance of AI systems, in particular frameworks and toolkits developed and published by RAi UK-funded projects and by other organisations.

We examine evidence of how and why these tools are used, and of their usefulness and their limitations. We look for lessons that could improve AI assurance in the future, including potentially using AI-based tools to support AI assurance.

The aim is to identify priority areas and key questions for RAi UK and other researchers and organisations to explore further, with the goal of improving the fit between the supply of, and demand for, tools that support organisational assurance of responsible AI.

Responsible Ai UK and our partner for this event, Confiance.ai, brought people together from across government, public services, and industry, as well as project teams that develop support for responsible AI. The aim was to clarify how people should approach the development and use of toolkits and/or frameworks in practice. We wanted to synthesise findings from our research across the programme and to find and address any gaps.

The workshop took place online on Thursday 10 April 2025.

RAi UK project presentations:

- **RAKE:** Responsible Innovation Advantage in Knowledge Exchange
- **Transparency Regulation Toolkits**

Additional presentations from industry and government:

- **Confiance.ai** program, France
- **Fraunhofer Institute** for Intelligent Analysis and Information Systems, Germany
- **DSIT** (Department for Science, Innovation and Technology), UK

After the presentations, attendees broke out into groups to reflect and discuss.

Introduction: Frameworks and Toolkits for Responsible AI

AI toolkits are collections of software, resources, and guidance designed to help individuals and organisations develop, deploy, and manage AI systems responsibly and effectively.

These toolkits can take many forms and address various aspects of the AI system lifecycle, from discovery and planning, through to deployment and auditing, including responsible AI practices, data curation and governance, model training and evaluation, and specific AI applications. They can be used to inform decision-making and implementation about how to develop, apply, or deploy AI tools within an organisation; and they can be used as a tool to demonstrate compliance with internal policies (of both AI system developers and users), contractual commitments, external technical standards, and legislative and regulatory requirements.

The landscape of AI governance is becoming increasingly filled with a plethora of toolkits, frameworks, and guidance documents that promise to deliver "responsible," "ethical," or "trustworthy" AI. Yet as these collections multiply, we face a growing paradox: more tools and frameworks do not necessarily mean an accelerated adoption of more responsible AI or even more adoption of toolkits. There is a lack of evidence around how these are being used, what works well or less so (including by sector or use case), the fit between supply and demand, and whether these tools continue to be effective over time as the capability of AI technologies and the range of applications for them both continue to grow. The proliferation of AI toolkits risks creating a side industry of superficial compliance where organisations can claim responsibility without meaningful change. There's a risk that each new framework may create a sense of progress, even though some core issues still need attention.

Looking at toolkits and frameworks produced by academia, industry, and governments, it is possible to identify a few characteristics that are present in a significant number of them, as follows:

1. Process-heavy solutions that focus on traditional documentation and compliance practices.

2. Broad principles without contextual nuance, making practical implementation challenging, considering that the way principles are understood and interpreted may vary (see below for the exploration of transparency in the RAKE project).

3. False sense of security through "checkbox compliance".

4. Failure to address issues across the AI lifecycle (e.g. considering unintended consequences at design time or continuous monitoring beyond deployment, or assurance of procured services).

5. Lack meaningful metrics and references to certification bodies/technical authorities for measuring success beyond self-assessment.

Do we need new frameworks and toolkits?

The UK Government has consistently expressed its view that AI assurance and compliance can support confidence in adoption of AI and thus help to drive wider productivity in the economy.[1] The underlying presupposition is that AI is insufficiently reliable or useful without assurance tools, and that compliance can be attained without substantive legislative change.[2] Current toolkits operate under the assumption that existing organisational structures and incentives cannot accommodate ethical AI considerations and therefore require supplementation. This fundamental premise deserves questioning.

The OECD has reviewed "trustworthy AI" toolkits and frameworks, publishing a catalogue of over 1000 such tools, though without assessments of their relevance or effectiveness. There is little evidence that any of these are widely used; and limited evidence of business benefit or resulting accelerated adoption of AI.[3]

A UK Government announcement in November 2024 counted 524 firms supplying AI assurance goods and services in the UK, and 84 UK-based specialised AI assurance companies.[4] That was reported to be a bigger market than in the US, Germany, and France, but the government's assessment was that it was still below the full size of the potential UK market for tools for assurance and compliance relating to AI.[5] Crucially, it was assessed that 'a lack of understanding among consumers of AI assurance about the risks posed by AI, relevant regulatory requirements, and the value of AI assurance is currently limiting demand for AI assurance tools and services.'[6]

A truly effective approach would require organisations to confront uncomfortable
fundamental questions about who benefits from AI systems, who bears the risks, and whether certain applications should be pursued at all – it is rare for toolkits to ask existential questions about 'whether' to use AI and more usual to focus on 'how' to deliver AI solutions. At this 'delivery' end, there is not yet a good answer about whether there is a way to build such toolkits to achieve meaningful business benefits that organisations can quantify and therefore invest time and money to achieve. A clearly communicated strategy for AI regulation, supported by a clear framework of laws, regulation, guidance, best practice, and technical standards would also serve to reduce the ambiguity fuelling inaction.

RAi UK has funded projects that are developing and publishing tools in this space. We want to develop a better-informed view of what these tools are for, who wants them and uses them, their relevance and utility to users, and how they can fit into plans to adopt AI, from single organisations up to the national level.

We want to know how these are built, what kind of difference they can make, whether they are driven more by insight or empirical evidence, and whether that can be measured. We aim to present useful insights to users, practitioners, and policymakers.

We use a relatively broad concept of assurance of responsible AI, because it has different meanings to different professions and in different industries, for instance, in engineering compared with procurement.

We asked these questions:

What drives the lack of adoption? Is it simply a lack of awareness – at least in certain sectors – of the existence of these frameworks, to which the solution would just be a matter of promoting existing toolkits and frameworks? Or do these toolkits fail to meet the actual needs of potential users? Could the issue lie in the absence of internal mechanisms, processes, and procedures within businesses to train staff, test tools, and oversee their application? Is the rapid pace of advances in AI development, combined with uncertainty in the regulatory landscape, making businesses hesitant to invest in or adopt these toolkits? Is it because the intended users of these toolkits are not involved in their design, development, or testing? Or is it due to a lack of empirical evidence demonstrating the value, impact, or practical usefulness of these toolkits?

Are current tools demand-driven? What feedback mechanisms exist to ensure relevance and usability?

What organisation types are missing in the ecosystem that would enable understanding both the need for, and deployment of, these tools? Where should responsibility lie in leading this?

[1] Department for Science, Innovation, and Technology. 'Guidance: Introduction to AI assurance' (UK Gov, 12 February 2024) https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance; Department for Science, Innovation, and Technology. 'Assuring a responsible future for AI' (UK Gov, 05 November 2024) https://assets.publishing.service.gov.uk/media/672a2ca440f7da695c921b7c/Assuring_a_Responsible_Future_for_AI.pdf; Department for Science, Innovation, and Technology & Government Digital Service. 'Artificial Intelligence Playbook for the UK Government' (UK Gov, February 2025) https://assets.publishing.service.gov.uk/media/67aca2f7e400ae62338324bd/AI_Playbook_for_the_UK_Government__12_02_.pdf > All accessed on 31 July 2025.

[2] The preferred option for the UK Government is 'to delegate' [AI regulation] to existing regulators with a duty to regard the principles, supported by central AI regulatory functions . Existing regulators have a 'duty to have due regard' to the cross-sectoral AI governance principles, supported by central AI regulatory functions [with] no new mandatory obligations for businesses. Department for Science, Innovation and Technology. (2023). A pro-innovation approach to AI regulation (CP 815). His Majesty's Stationery Office. https://assets.publishing.service.gov.uk/media/64cb71a547915a00142a91c4/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf> Accessed 31 July 2025

[3] As of 31 July 2025, the list now stands at 939 with a number of frameworks being withdrawn at the last biannual review. OECD Catalogue of Tools & Metrics for Trustworthy AI. - https://oecd.ai/en/catalogue/tools

[4] Department for Science, Innovation, and Technology. 'Assuring a responsible future for AI' (UK Gov, 05 November 2024) https://assets.publishing.service.gov.uk/media/672a2ca440f7da695c921b7c/Assuring_a_Responsible_Future_for_AI.pdf> Accessed 31 July 2025

[5] *Ibid.*

[6] *Ibid.*

Presentation Summaries

RAKE: Responsible Innovation Advantage in Knowledge Exchange

A collaboration between the Universities of Nottingham, Oxford, and Warwick, with Kainos as the main industry partner.

RAKE is an Impact Accelerator project funded by RAi UK, looking at the adoption and use of responsible innovation (RI) approaches within the whole AI innovation chain. This project has been investigating how RI is and could be better embedded within Centres for Doctoral Training and businesses, to support a new generation of RI into practice and improve responsible AI development and deployment. Challenges and the blockers, as well as opportunities and benefits, have been identified. In particular, RAKE has explored with the business sector the use of an RI tool: Responsible Innovation Prompts and Practice Cards (RI Cards).[7] This high-level tool was developed as part of a previous research project within Horizon Digital Economy Research, co-created with stakeholders, and refined through the TAS programme at the University of Nottingham (both UKRI-funded programmes). The RI Cards are designed to highlight 16 different aspects of RI, each of which poses key questions and prompts for actions to consider.[8] Initially, the project engaged with delivery managers from Kainos and assessed pre-existing knowledge of RI and responsible approaches within their work. Then, project members introduced the RI Cards to Kainos' delivery managers, who then used the RI cards as a support tool within their projects. RAKE followed up with a weekly survey for eight weeks to understand how the delivery managers were able to work with those cards within their teams and clients (Phase 1).

In general, the RI cards were welcomed and positively viewed. The responsible innovation approach is successful in prompting questions about whether a company should be doing something in a particular way, and whether AI is the right tool for a particular use. The physical process of handling the cards helped people to engage, more than having points on a board or thinking about the questions in the abstract.

There were blockers around extra time and uncertainty about how some companies' clients might respond. However, people found it very helpful to use the RI Cards as a framework to help them think about what the key questions and considerations for a context were. Delivery managers found this tool helped their teams to feel empowered to innovate in the "right way", providing structure, encouraging reflection, decision making, and creation of an RI action plan.[9] Interestingly, it was difficult to define all relevant stakeholders.

RI card activity during a RAKE workshop with Kainos, photo credit: RAKE Impact Accelerator

Kainos have found that a responsible innovation approach can be a business benefit: it may not speed things up, but it can make for better overall outcomes. At the time of writing this paper (Phase 2), RAKE is working with Kainos to incorporate the RI Cards tool – and learnings from Phase 1 - to develop an open-access actionable toolkit to support the AI business sector to embed RI into the end-to-end delivery life cycle. It aims to understand what the opportunity is for embedding specific aspects (and questions) of RI in the right part of the delivery life cycle and then to try and work out from that what mitigations or considerations to build into the process.

Transparency Regulation Toolkits for Responsible AI

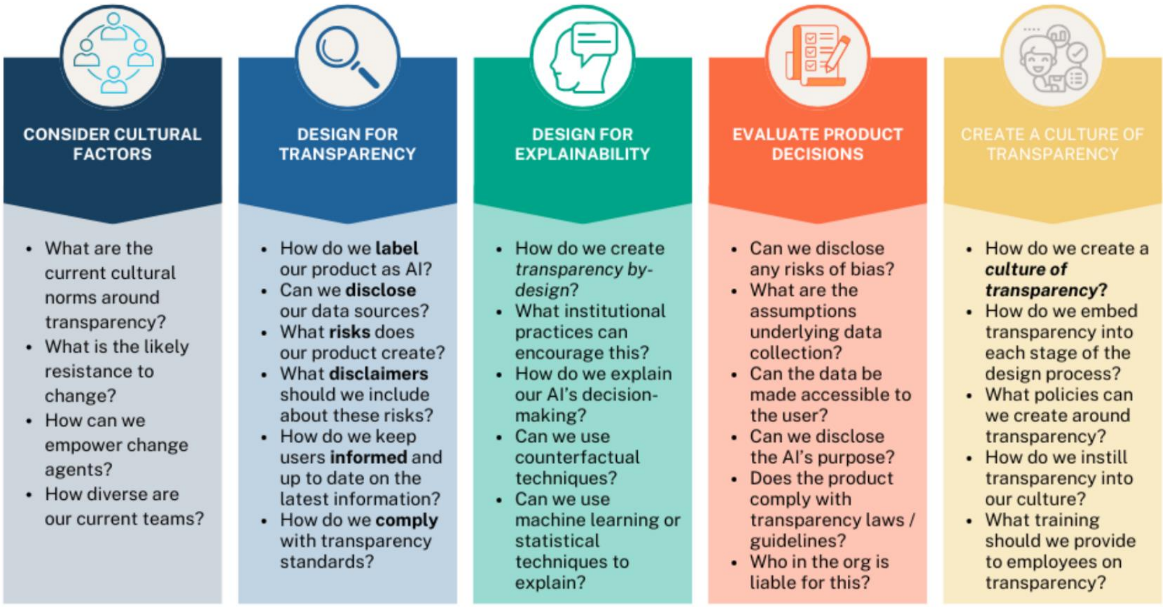
A collaboration between University of Bristol, University of Antwerp, scholars, lawyers, and Small and Medium-sized Enterprises (SMEs).

Despite the UK and the EU having different requirements, it became apparent early on that UK companies needed to be compliant with EU regulation and requirements for AI. It was therefore helpful for organisations in the EU to understand UK regulation, so the planned two toolkits became one. The target market was SMEs with fewer than 250 employees and an annual turnover of less than £50 million or the equivalent in euros.

The view was that these SMEs lag behind larger enterprises in AI adoption, in part because of the cost of compliance, IT infrastructure, and expertise to manage the regulatory dimensions, and partly because of the knock-on impacts in terms of risk appetite and capacity. There has been research on how large enterprises manage this, but less on smaller enterprises. The programme aimed to close this gap by providing something affordable, practical, and accessible on how to meet transparency requirements.

They identified some core dilemmas. Firstly, granularity: the desire to make the toolkit straightforward clashed with the complex and often ambiguous realities of transparency. There were many different ways of being transparent, and different goals and different audiences for which transparency was sought. Second, the regulatory and technological landscape was – and continues to be – in constant flux, with evolving AI tools and shifting legal requirements.

The team addressed the granularity problem with a layered approach. The toolkit itself offered accessible, high-level guidance through questions and recommended actions. Beneath this sat an interactive report, providing links to further tools and in-depth guidance. To tackle ambiguity, a reflection stage was built into the toolkit, encouraging SMEs to think critically about their objectives, intended audiences (e.g. customers, users, employees, or regulators), and what transparency means in their specific context.



CONSIDER CULTURAL FACTORS

- What are the current cultural norms around transparency?
- What is the likely resistance to change?
- How can we empower change agents?
- How diverse are our current teams?

DESIGN FOR TRANSPARENCY

- How do we **label** our product as AI?
- Can we **disclose** our data sources?
- What **risks** does our product create?
- What **disclaimers** should we include about these risks?
- How do we keep users **informed** and up to date on the latest information?
- How do we **comply** with transparency standards?

DESIGN FOR EXPLAINABILITY

- How do we create *transparency by-design?*
- What institutional practices can encourage this?
- How do we explain our AI's decision-making?
- Can we use counterfactual techniques?
- Can we use machine learning or statistical techniques to explain?

EVALUATE PRODUCT DECISIONS

- Can we disclose any risks of bias?
- What are the assumptions underlying data collection?
- Can the data be made accessible to the user?
- Can we disclose the AI's purpose?
- Does the product comply with transparency laws / guidelines?
- Who in the org is liable for this?

CREATE A CULTURE OF TRANSPARENCY

- How do we create a ***culture of transparency?***
- How do we embed transparency into each stage of the design process?
- What policies can we create around transparency?
- How do we instill transparency into our culture?
- What training should we provide to employees on transparency?

Governance and Cultural Change: Extract from Krook J, Winter P, Blockx J, and Downer J, (2024) 'Transparency Regulation Toolkit for Responsible AI,' Responsible AI UK, International Partnership.

The programme found that transparency means very different things across sectors and organisational settings. The team took a pragmatic approach, focusing on best practices for legal and regulatory compliance, while also giving organisations tools to define and implement transparency in their own terms.

The issue of regulatory flux remained a challenge, as the team lacked the capacity to update the toolkit continuously. They approached potential strategic partners interested in maintaining the toolkit or evolving it into a more dynamic solution – specifically, an AI software that users could interact with. This led helpfully to the question of whether interactive AI systems could build on and replace these toolkits. An offshoot of the programme is creating a ChatGPT type of AI transparency platform, and they propose to test this with 20-30 SMEs, so it provides each company with tailored support, adapting the guidance in the toolkits to suit each business's specific needs and sector (pending funding). Another emerging idea is to develop a second software tool: an AI Best Practice Development Platform for SMEs. This would be built in collaboration with Oldfield Consultancy and The Institute of Science and Technology (IST), subject to funding.

Confiance.ai

There are many papers on ethics, but for the European Union, the high-level Expert Group on AI did a lot of work developing the ethics that are implemented into regulations and will have some longevity over 10 or 20 years, and in harmonised standards that might be revised after five years.

The goal of Confiance.ai was to address people working on application engineering for trustworthy critical systems, distributed AI, and generative AI. The programme has looked at all aspects of engineering, (data engineering, knowledge engineering systems, safety, human factors engineering) and addressed several industries, across critical sectors.

The programme produced a methodology for audit and created a catalogue of tools and use cases. The tools have different levels of maturity, but they have been tested on at least one use case. Confiance.ai is also involved in the AI Trust Alliance, which works on indicators, evidence, and has a scoring chain, which will be used to create an AI Trust Alliance label.

The overall objective is to prepare organisations to be fully compliant with standards: National, European, and international. RAi UK has formed a partnership with the European Trustworthy AI Association, the follow-on activity from Confiance.ai which will maintain the catalogue and all these tools.

Fraunhofer Institute

Fraunhofer leads a programme on certified AI in Germany, developing and testing principles and tools to assess the trustworthiness of AI systems. The output, produced three years ago, was an assessment catalogue with a risk-based approach to evaluate the trustworthiness of AI systems. Since then, there have been some assessments based on this catalogue and the created decision records, a concept that originally comes from software development. These provide guidance for development teams to make decisions about trustworthy AI and to document those decisions. The programme has also worked with a small number of companies to build AI governance.

The programme developed a process for semantic testing or weak slice discovery to explore the weaknesses of neural networks. This inserts a semantic structure in the input space and systematically searches for weaknesses within that space to identify the conditions that can create weaknesses and address them. This is just one example. The programme explored how to integrate tools to measure bias in datasets, and how to develop a vision for a multi-party environment that can integrate tools from different parties and stakeholders. The programme has published an architecture of what this type of platform could look like, how different roles that are using this platform are defined, and how to automate corresponding workflows. It also developed utilities that allow you to onboard new tools on this platform.

The programme implemented testing and assessment workflows for different examples, object detection, and error testing on the platform.

Companies implementing the AI Act are concerned about how, in practice, they can incorporate it into their development teams. There are discussions with companies building up business models with third party assessments. Some companies are also thinking about how to automate compliance with the AI Act.

Department for Science, Innovation and Technology (DSIT): the UK AI Opportunities Action Plan

The UK AI Opportunities Action Plan ("The Action Plan") is the bedrock of government's approach to AI regulation and its use and adoption across the country. In government's response to The Action Plan, explicit mention is made of government's intention to 'support DSIT's existing programme of work designed to stimulate the AI Assurance ecosystem' – as articulated in Assuring a responsible future for AI.

Relevant to responsible AI, there are three pillars: awareness, capability, and governance, with an objective that senior leaders become intelligent customers of AI products and services, aware of their obligations to develop and implement responsible AI systems which are compliant with regulation, and embrace design and implementation principles, reflecting the wider ambitions of the AI growth strategy.

Toolkits are intended to help companies develop and use frameworks for assurance and risk management. This is significant, as the government is taking a light approach to regulation through legislation at this stage. A key part of enabling responsible adoption is to have assurance in place to understand the risks in terms of privacy and fundamental rights, and there is a proper regulatory framework in place for companies to understand and meet obligations.
That includes ensuring that there are internationally aligned standards, so the UK will focus heavily on working alongside technical standards bodies including BSI, ISO, IEC, and ETSI. The aim is to have sector-specific rules and guidance, rather than explicit AI regulation across the whole AI sector.

Department for Science, Innovation & Technology

Assuring a responsible future for AI

Accelerating the growth of the UK's AI assurance market

Cover image: DSIT, Assuring a responsible future for AI, published 6 November 2024

In this context, the government is seeking to create its AI assurance platform as a one stop shop for information and resources, in particular to support scale ups and SMEs in becoming intelligent customers fully aware of the regulatory environment in which they are working, understanding the regulations, the technical standards, and the expectations that are placed on them. This should reduce some of the burden of identifying compliance requirements and support ongoing dialogue between the regulators and the private sector in order to understand mutual obligations and how those will be fulfilled.

The government plans to support organisations to establish robust management practices for AI systems, ensuring that they can have ethical and effective AI integration into their operations and allowing them to navigate the complexities of AI deployment and address potential risks. The government will create a framework in which organisations can evaluate and enhance their AI management systems and processes. It is focusing on assessing the organisational processes rather than on the AI products or services themselves to promote responsible AI development. The tool provides a structured questionnaire that reviews essential AI governance components. AI Management Essentials will provide the baseline standard for the management of organisations developing and deploying their products and services, aligned with existing international technical standards.

Top: Virginia Portillo (RAKE) and bottom: Peter Winter (Transparency Regulation Toolkits), presenting at the RAi UK Impact Accelerator and International Partnership Networking Event, Royal Society, London, March 2025

7. Portillo, V., Greenhalgh, C., Craigon, P. J., & Ten Holter, C. (2023, July). Responsible Research and Innovation (RRI) prompts and practice cards: A tool to support responsible practice. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems* (pp. 1-4).

8. https://www.nottingham.ac.uk/research/groups/mixedrealitylab/projects/ri-cards.aspx >Accessed 31/07/2025

9. https://rai.ac.uk/responsible-innovation-into-action-rake-engagement-with-the-business-sector/ >Accessed 31/07/2025

Prof Sarvapali (Gopal) Ramchurn, RAi UK CEO and Nicolas Rebierre, General Manager European Trustworthy AI Association, Royal Academy of Engineering, March 2025

Exploring the use of Toolkits and Frameworks

1. Adoption

What drives the lack of adoption? – Is the solution just a question of promoting the toolkits and frameworks, or do these toolkits not meet the needs of potential users?

The discussions surfaced interacting factors affecting the uptake and use of these frameworks and toolkits.

Incentives and regulations (carrots and sticks): Unless they see a clear requirement for specific compliance with regulation, or clear challenges to reputation or brand from using a novel technology application, companies tend to keep to existing procedures.[10] Legislation tends to relate to outcomes rather than underlying technologies, so it may be difficult to understand what measures will deliver compliance, and how a toolkit can help meet that need. Overly general aspirations for responsible use of AI may fail to translate into meaningful drivers for companies and perhaps whole sectors.

The fast evolution of technology, laws, and regulations means toolkits may quickly become out of date, or potential users fear they will become out of date. They will be reluctant to commit resources to them, if future developments in regulation might demand something different or more specific (for instance, the AI Act's harmonised standards are not ready).[11]

Enforcement and visible lack of it are also relevant. If companies do not see legislation being enforced on organisations or activities like theirs, they will experience less pressure.[12]

There is insufficient empirical evidence of the effectiveness, cost, and benefits of frameworks, especially for SMEs. In addition, it may not be clear who the increased transparency aims to serve.

The mentality of "move fast and break things" is celebrated among some technology professionals, who may, as a result, not prioritise responsible practice or seek to use tools to assure it.[13] Most start-ups fail, and founding and running a start-up typically involves an appetite for and expectation of risk. Risk of causing harms or of falling foul of regulators may be accepted with other risks by some technology sectors, companies, and individuals through the early stages of growth. This stance may be encouraged by investors in spinouts and start-ups, who do not see an incentive in supporting innovators to adopt responsible toolkits.

Guidance that explicitly links the use of new applications to existing compliance obligations might provide a stronger incentive. Certification could encourage good practice if it is clear how it can help reduce liabilities.[14] Leadership for assuring responsible AI implementation is key in all sectors working in AI. Responsible AI leadership programmes should be encouraged and fostered within the academic, public, and industry sectors to ensure responsible development, deployment, decommissioning, and governance of AI. An internal team of driven Responsible AI leaders within an organisation/programme/project can drive positive change, influence the implementation of responsible AI using and adapt tools and existing frameworks to support Responsible AI practice. This would also ensure the sustainability of practice and bespoke implementation according to specific areas of work.[15]

What do corporate insurers require from companies they provide cover for and that use AI applications? It could be enlightening to know if insurers are developing requirements and what those are.

Choosing a solution: Users may not know where to find existing resources, free or paid-for, or how to choose one. There are many available toolkits, for apparently similar objectives. This is not a mature market with norms and expectations, and users can lack confidence when selecting solutions. AI is now a very broad and overused term; therefore, “AI risks” may cease to be a helpful term. Exploration of new risks and requirements in specification application areas may well need more specific terms.

In this respect, senior leaders in private and public sector organisations may not know what questions to ask, where to look, or what to look for, regarding toolkits, leading them to muddle through or rely on traditional methods. Expectations and obligations for responsible use of AI are not yet normalised parts of organisational practice.

While OECD provides tools, there is a lack of standardised metrics for evaluating what these tools do. Users may not see clear, demonstrable value from using them. Case studies showing “better outcomes” are needed. Users may not fully know what they want from a solution: there appears to be demand for both one-stop solutions and highly bespoke ones. More case studies could help educate both demand and supply.

Use and usability: Users need tools that are responsive to their specific contexts. Tools may not be tailored enough to specific needs or may lack sufficient guidance for real-world application. Developers may not have detailed knowledge of users’ practical needs.

The ease with which RI tools can be integrated into project management tools already in use by AI systems developers also needs to be considered. Users may not know how, where, or when to integrate tools into existing workflows, AI lifecycles, and systems. There can be challenges with transparency in AI supply chains (confidentiality, trade secrets, opacity), which can complicate the use of a toolkit.

Use requires endorsement by senior staff, but it may also involve additional training for them as well as for operational staff.

Much of this is not new. We know that technology adoption can fail due to insufficient support, monitoring, and assurance from organisational structures and cultural norms. Adoption of toolkits will similarly be undermined where it is not well supported by leadership, management, and dedicated training to embed use and maintain motivation. To deliver benefits and confidence, toolkits need to be built into everyday organisational practices, which involves establishing organisational mechanisms to support routine use, including those that will enable routine auditing and governance. It may not be clear whose responsibility it is to maintain the use of a framework and apply any learning from it.

SMEs particularly struggle with constraints on time, finance, and expertise for training, accessing appropriate best practice guidance, and implementation. It can be difficult for an SME to make a decisive business case: implementing an assurance toolkit can seem a disproportionate cost if the value of the regulatory compliance delivered is difficult to estimate. Benefits should ideally improve with integration into practices and then tracking of use and outcomes. Those all incur ongoing costs.

Using a framework and/or toolkits without embedding them as part of a process (and in a systematic way) can become a box-ticking exercise:[16] a formality that does not deliver results, similar to “"greenwashing,”" without genuine integration and understanding and very likely failing to provide benefits either in a better understanding of possible impacts or of compliance.[17] Evaluations can start too late in development, making changes discovered during assessment expensive.

Really effective assurance of a corporate use of AI might need to involve a broad range of stakeholders in design and use. Currently, organisations may not understand this need, and many lack capacity or the right relationships to convene those stakeholders. Therefore, it may be valuable to find ways to help organisations understand the need for broader and more inclusive assurance processes.

2. Design and feedback
Are these tools demand-driven, and what feedback mechanisms exist?

Drivers for creating toolkits: There is a mixed picture, but most current toolkits are not truly demand-driven and, unlike the ones described in this whitepaper as part of RAi UK (i.e., RAKE and Transparency Regulation Toolkits for Responsible AI projects), are often developed without much input from real users. Development is often driven primarily by efforts to advance responsible AI.

There are real challenges for developers of toolkits in understanding demand so as to match responsible AI actions to deliver against specific outcomes or goals within disciplines and organisations.

Requirements voiced by potential users can be vague, forcing developers to guess what users really need. It can be difficult to express demand for responsible AI, because it has so many facets in different fields. AI tools and the potential range of applications are developing rapidly, so it is difficult for assurance toolkits to match contexts. Complex ecosystems make it unclear who values these tools.

The variety of options and the difficulty of matching a tool with particular needs can lead to decision paralysis. This may be less the case where tools are developed to respond to specific industry needs (transparency in finance, ethical considerations in healthcare).

There also appears to be activity contributing to this space, and users are following others within their sector. Demand varies a lot between sectors. There are some expectations: for instance, international (IEEE , ISO) certification (e.g., Institute of Electrical and Electronics Engineers and International Organisation for Standardisation) is perceived as necessary despite being seen by some as cumbersome and expensive. Moreover, SMEs often lack the capacity to absorb such overheads.

Feedback: The mechanisms for gathering and, more importantly, acting upon user feedback are largely underdeveloped, lacking transparency, and not systematically integrated, contributing significantly to the ongoing relevance and usability challenges. There is little transparency about how user input influences the development of tools.
Overall, mechanisms to assess and ensure relevance and usability appear to be limited and insufficient. Also, there is not a universal mechanism to track the impact of implementing these tools, as there is no one-size-fits-all approach to do this.

Even when users provide feedback, it is not always clear how this is (or might be) used to make improvements. There is a lack of capacity to convene and engage the breadth of stakeholders involved in developing and using these systems. There can be an excessively narrow view of the value of feedback. A “whole ecosystem view” is needed. Existing feedback can be siloed or focused on proprietary models, not broader improvement to toolkits. It is not clear which organisations and mechanisms could be most appropriate to collect and use feedback and be trusted.

Possible solutions: Bodies that drive standards in sectors and professions could be well positioned to provide feedback on relevance and usability. There may be lessons to learn from sectors like healthcare with long-established ethical and assurance practices.

Transparency about how feedback is used would encourage active participation and build trust. It was suggested that users would be more engaged if they could see their feedback being acted on, as this visibility helps build trust and encourages continued use.

Companies can create comprehensive feedback mechanisms, but are mostly concerned with their own models and competitive practices. We need a better understanding of current practices to integrate by co-creating with users' practices. Maturity models need to be well-defined if they are to be successful in enabling organisations to develop standards and repeatable implementation processes.

3. What organisation types are missing in the ecosystem that would enable understanding both the need for and deployment of these tools?

There is no global or overarching governance or certification body, and leaving it to existing regulators leaves a fragmented landscape. Some form of certification/standardisation authority may be valuable, along the lines of the National Physical Laboratory, British Standards Institution, and United Kingdom Accreditation Service.

There are several roles to be delivered. More research on setting up standards on responsible AI metrics could identify appropriate levels of specificity to sectors. The RAi UK Keystone project AdSoLve is pursuing this objective, using law and medicine use cases to identify requirements for benchmarks that are capable of assessing the performance of LLMs in real-world settings. There is demand for providers of advice on development and adoption, audit and certification. There is a role for organisations that "assure who assures", providing legitimacy and accreditation for designers and providers of tools, and performing post-market surveillance as Yellow Card does for medical devices.

Assurance of responsible AI would also be improved by the participation of civil society organisations representing the full range of relevant user and societal perspectives. Representatives of users (and subjects) of services and consumer advocacy groups could help to ensure that responsible AI tools address real-world issues that affect consumers. Community organisations can provide insights into how AI impacts vulnerable populations and suggest ways to mitigate negative effects.

Unions and industry-specific professional bodies could gather useful, informative, unguarded views on the use of tools from employees.

Accreditation bodies and authorities, standardisation institutes, and organisations from the TIC sector (Testing, Inspection, and Certification) have relevant expertise and reach, including those that provide certification of skills in software engineering and in data science and setting curricula for skills, like the British Computer Society and the Institute of Electrical and Electronics Engineers.

AI Offices might best deliver some roles within existing domain-specific bodies, which could provide context-specific guidance rather than generic information.

Organisations providing oversight in finance could provide assurance to investors on reputability, quality of the product, and investment guidelines based on good practice for startups and scale-ups, effectively driving adoption through financial incentives. There is also a need for guidance on the inherently multi-disciplinary and participatory nature of the successful deployment of assurance tools.

Organisations representing key professional users of AI systems in healthcare, social work, law, finance, and civil servants, have traditionally played an important role in governing professional standards.

There are many potential roles here responding to sectors and contexts within an overall ecosystem of collective responsibility.

Government and regulators can provide stronger drivers with statutory standards and mandatory frameworks. They need to contribute to how tools are developed. The Information Commissioner's Office was mentioned as a good example of a sector-specific regulator offering tools and support.[18] Government procurement is also an instrument to encourage framework adoption. Still, it can bring additional burdens to the process, turning it even more complex and bureaucratic, depending on how “encouragement” is implemented, a point that should be considered especially in light of SMEs' participation in the bidding competition.[19] International coordination on certification would add force and recognition.

Professional bodies and industry organisations could be influential in spreading learning systematically, driving standards for those professions, as feedback conduits, and for facilitating network effects and sharing of best practices. For example, in 2024, the UK Law Society released a guide on the use of Generative AI for law practitioners.[20]

Under the proper incentives (e.g. government recognition, rewarding, and positive brand publicity), major technology companies could leverage their resources and expertise to help smaller companies understand and meet technical guardrails.

In more sensitive application areas, there should be mechanisms to involve those directly and indirectly affected by AI applications in the lifecycle for assurance tools, establishing shared responsibility for defining what is needed.

[10] The relationship between regulations, compliance and reputational risk has been previously addressed in the following works within the environmental context: Prakash, Aseem. Greening the Firm: The Politics of Corporate Environmentalism (Cambridge University Press 2000); Gunningham, Neil, Kagan, Robert A. and Thornton, Dorothy. Shades of Green: Business, Regulation and the Environment (Stanford University Press, 2003); Mehta, Alex and Hawkins, Keith. ‘Integrated Pollution Control and Its Impact: Perspectives from Industry’ (1998) 10 Journal of Environmental Law 61–77. Robert A. Kagan, Neil Gunningham and Dorothy Thornton. ‘Explaining Corporate Environmental Performance: How Does Regulation Matter?’ (2003) Law and 37 Society Review 51–90.

[11] The reluctance to commit resources to toolkits if future regulations demand different parameters can be explained by the notion of regulatory uncertainty as a compliance cost. See: Cordes, J. J., Dudley, S. E., & Washington, L. Q. ‘Regulatory compliance burdens.’ (2022) George Washington University, Regulatory Studies Center. Available at: https://regulatorystudies.columbian.gwu.edu/sites/g/files/zaxdzs4751/files/2022-10/regulatory_compliance_burdens_litreview_synthesis_finalweb.pdf> Accessed on 31 July 2025.

12. Regulatory and compliance theory studies have investigated that idea, showing the corelations between enforcement measures and pressure to abide by the regulations. See: Thornton, Dorothy, Kagan, Robert A. and Gunningham, Neil. 'General Deterrence and Corporate Environmental Behavior' (2005) 27 *Law and Policy* 262–88; Gunningham, Neil, Thornton, Dorothy and Kagan, Robert A. 'Motivating Management: Corporate Compliance in Environmental Protection' (2005) 27(2) *Law and Policy* 89–316; Mendeloff, John and Gray, Wayne. 'Inside OSHA's Black Box: What is the Link Between Inspections, Citations and Reductions in Different Injury Types?' (2004) 27 *Law and Policy* 219-37; Shimshack, Jay and Ward, Michael. 'Regulator Reputation, Enforcement, and Environmental Compliance' (2005) 50 *Journal of Environmental Economics and Management* 519-40; Kazumasu Aoki and John W. Cioffi. 'Poles Apart: Industrial Waste Management Regulation and Enforcement in the United States and Japan' (1999) 21 *Law and Policy* 213-45; Financial Conduct Authority. 'Behaviour and Compliance in Organisations' (2016) Available at: https://www.fca.org.uk/publication/occasional-papers/op16-24.pdf> Accessed on 31 July 2025.

13. About the "Move fast break thing mentality", see: Birkinshaw, Julian. 'Move fast and break things: Reassessing IB research in the light of the digital revolution' (2022) 12 *Global Strategy Journal* 619-631. 10.1002/gsj.1427; John RR, 'Move Fast and Break Things: How Facebook, Google, and Amazon Cornered Culture and Undermined Democracy By Jonathan Taplin' (2018) 92(1) *Business History Review* 191-193. doi:10.1017/S000768051800020X.

14. About the need for clarifying the role of certification mechanisms in liability regimes, see: Hanna Schebesta, 'Risk Regulation Through Liability Allocation: Transnational Product Liability and the Role of Certification' (2017) 42(2) *Air and Space Law* 107-136 https://doi.org/10.54648/aila2017011; Boehm, T. C., & Ulmer, J. M. 'Product Liability: Beyond Loss Control–An Argument for Quality Assurance' (2008) 15(2) *Quality Management Journal* 7-19. https://doi.org/10.1080/10686967.2008.11918063.

15. About the role of leadership in driving organisational change, see: Errida A, Lotfi B. 'The determinants of organizational change management success: Literature review and case study' (2021) 13 *International Journal of Engineering Business Management*. doi:10.1177/18479790211016273.

16. For a better understanding of ticking the box exercises and their risks, see: Van Vuuren, H. J. 'The disclosure of corporate governance: a Tick-Box exercise or not?' (2020) 12(1) *International Journal of Business and Management Studies* 50-65; Reddy, Bobby V. 'Thinking Outside the Box–Eliminating the Perniciousness of Box-Ticking in the New Corporate Governance Code' (2019) *Modern Law Review* 692-726. https://doi.org/10.1111/1468-2230.12415.

17. About greenwashing, see: Mutua K, Powell-Turner J, Spiers M, Callaghan J. 'An In-Depth Analysis of Barriers to Corporate Sustainability' (2025) 15(5) *Administrative Sciences* 161. https://doi.org/10.3390/admsci15050161; Free, C., Jones, S. and Tremblay, M.-S. 'Greenwashing and sustainability assurance: a review and call for future research' (2024) *Journal of Accounting Literature*. https://doi.org/10.1108/JAL-11-2023-0201.

18. For example: ICO. Artificial Intelligence guidance. Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/> Accessed on 31 July 2025.

19. About the role of procurement in setting standards and practices related to policy goals, see: Howe, J. The regulatory impact of using public procurement to promote better labour standards in corporate supply chains. In Fair Trade, Corporate Accountability and Beyond (Routledge, 2016); Oishee Kundu, Elvira Uyarra, Raquel Ortega-Argiles, Mayra M Tirado, Tasos Kitsos, Pei-Yu Yuan, 'Impacts of policy-driven public procurement: a methodological review' (2025) 52(1) *Science and Public Policy* 50–64, https://doi.org/10.1093/scipol/scae058; OECD, Public procurement can be used strategically to achieve broader policy goals. Available at: https://www.oecd.org/en/topics/public-procurement.html> Accessed on 31 July 2025.

20. Available at: https://www.lawsociety.org.uk/topics/ai-and-lawtech/generative-ai-the-essentials> Accessed on 31 July 2025.

Questions and Opportunities

There are evidently gaps between current practices and resources, and what would constitute widespread assurance of responsible and trustworthy application of AI.

There is a lack of empirical research on organisations using these toolkits, making it hard to make detailed assessments of how usable and effective they are, or of the relationship between the use of AI and the use of measures for responsible AI.

Assuring AI can potentially help people understand and meet their responsibilities, still, more needs to be done to show why and how using AI responsibly will be of value to organisations and their employees, including for compliance, transparency, and liability. There are questions about where agency and decision-making sit and should sit. At the moment, it is not clear which leadership roles are responsible for ensuring the responsible use of AI.

In some regulated industries, domain regulators have already issued some guidance on firms' internal organisation, governance structure, and decision-making responsibilities. Within such sectors, looking at those regulations can help organisations understand how leadership should be engaged and their responsibility for ensuring responsible AI adoption. In this context, regulatory agencies could add to their guideline specific guidance about AI. For example, in the banking and financial sector, where AI is used for transaction monitoring purposes, the Financial Conduct Authority's Financial Crime Guidance of 2025 (Section 3.2) has considered a poor practice for "senior management to have an unrealistic expectation of what automated monitoring systems are feasibly able to achieve".[21] Other agencies could follow the same example.

At this point, in many sectors, there have not been many public failures of corporate responsibility when using AI, which could encourage companies and their leaders to implement measures to avoid making similar mistakes. When this kind of public incident occurs, it might change views within a sector.

The sustainability of processes appears challenging. Improvements might be made by better use of feedback to improve toolkit design, fit, and usability, building stronger organisational capacity and culture, clarifying regulatory expectations and incentives, and enhancing communication and coordination across the ecosystem. Gathering better evidence about the use of toolkits in practice would require better monitoring.

In wider ecosystems, it is not clear who could consistently train the trainers, and audit the auditors, to improve the effectiveness, adoption, and deployment of tools for assuring responsible use. Given the importance of context to responsible use, sectoral approaches and responsible bodies might be effective.

Comparable forms of assurance: What learning can be captured and applied from drives to improve sustainability in industry? Sustainability missions have also combined appealing to ethics, responsibility to the full range of stakeholders, regulatory compliance, and business benefits. That has been a long journey which is not yet complete, but it should be possible to identify what has worked in terms of making assurance easier and more attractive to embed into business processes.

Automating assurance: We identified a major opportunity and challenge in the potential for automating tools for assuring responsible AI. Automation of assurance processes using AI might supersede toolkits and specialists. What could be the benefits and the risks? What kind of learning could be integrated, and how? Automation might well improve the consistency of some processes, but if automated assurance processes were trained on toolkits and frameworks that are relatively poorly based on evidence, faults could be multiplied. Automation that was too generic could become another box ticked rather than really addressing specific risks. What might be lost if assurance became less visible or invisible? There might be important differences between using AI to interrogate and navigate existing rules and regulations and using it to frame them.

[21] FCA, 'Financial Crime Guide: A firm's guide to countering financial crime risks (FCG)'. Available at: https://www.handbook.fca.org.uk/handbook/FCG.pdf> Accessed on 31 July 2025.

Further reading

Responsible Ai UK
https://rai.ac.uk/

AI Opportunities Action Plan, independent report for the Department for Science, Innovation and Technology (DSIT), UK
https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

Assuring a responsible future for AI, Department for Science, Innovation and Technology (DSIT), UK and The Rt Hon Peter Kyle MP
https://www.gov.uk/government/publications/assuring-a-responsible-future-for-ai

Confiance.ai
https://www.confiance.ai/
https://www.confiance.ai/foundation/

Fraunhofer Institute
https://www.fraunhofer.de/en/research/fraunhofer-strategic-research-fields/artificial-intelligence.html

Infosys
Responsible AI Toolkit, by Infosys, is an open-source solution that helps teams embed fairness, explainability, privacy, safety, and security directly into AI application. Its customizable and flexible design supports diverse AI models and deployment environments, reducing risks and enhancing trustworthiness in AI systems - https://github.com/Infosys/Infosys-Responsible-AI-Toolkit

Projects

RAKE: Responsible Innovation Advantage in Knowledge Exchange
Impact Accelerator
https://rai.ac.uk/new_projects/rake-responsible-innovation-advantage-in-knowledge-exchange/

Transparency Regulation Toolkits for Responsible Artificial Intelligence
International Partnership
https://rai.ac.uk/new_projects/transparency-regulation-toolkits-for-responsible-artificial-intelligence/

Contributors

RAi UK

Sarvapali (Gopal) Ramchurn
CEO RAi UK
University of Southampton

Pepita Barnard
University of Nottingham

Adarsh Valoor
University of Southampton

Aled Lloyd Owen
CoS, RAi UK
University of Southampton

Aislinn Bergin
University of Nottingham

Maria Waheed
University of Nottingham

Shuang Ao
University of Southampton

Isabela Parisio
King's College London

RAKE

Carolyn Ten Holter
University of Oxford

Virginia Portillo
University of Nottingham

Rob Procter
Warwick University

Paul Batterham
Kainos

Transparency Regulation Toolkits

John Downer
University of Bristol

Peter Winter
University of Bristol

Joshua Krook
University of Antwerp

Jan Blockx
University of Antwerp

Confiance.ai

Bertrand Braunschweig

Karla Quintero

Fraunhofer Institute

Maximilian Poretschkin

Daniel Becker

DSIT

Emily Campbell-Ratcliffe

Ben Hawes
University of Southampton
(Freelance)

Ashish Tewari
Infosys

Eliot Gillings
Royal Academy of Engineering

Patricia Shaw
Beyond Reach Consulting

Alec Thomas
Ministry of Defence

Reema Patel
Elgon Social

Paul Duncan
NPL

About Responsible Ai UK (RAi UK)

With a £35 million UKRI investment, RAi UK is a programme dedicated to delivering interdisciplinary research and fostering ecosystems, including international ecosystems, that support Responsible AI research and innovation. Through extensive consultations across the UK, RAi UK has identified emerging challenges in responsible AI and deployed over £17 million into projects aimed at accelerating the adoption of responsible AI practices and technologies. RAi UK brings research-based expertise that is connective, adaptive, and world-leading through field-building, and engagement with communities, publics, industries, and governments. The RAi UK research community includes expertise from across social sciences, law, engineering, computer science, and other disciplines, and aims both to achieve learning and to put it into practice and support that with new dedicated tools.

As well as informing our future work, we will create as much value as we can from projects we funded since the Programme's start in May 2023, in terms of evidence and practical policy ideas, for use by external policymakers, including government bodies, Non-Governmental Organisation (NGOs), and other international actors. We can share, develop, and promote enablers that can help AI work for everyone in different contexts internationally. RAi UK can take a leading role turning that into actionable knowledge and making it available globally to build toolkits and frameworks that people can use. We will also continue to act in a convening role, enabling new discussions and building networks with access to practical tools.

Contact

Email: info@rai.ac.uk
www.rai.ac.uk

RAi UK is supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1].