European Trustworthy AI Association

ADVANCING TRUSTWORTHY ARTIFICIAL INTELLIGENCE:

Lessons Learned and Emerging Challenges

White Paper 03, November 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*
- *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 paper examines the practical, policy, and technological dimensions of building and sustaining trustworthy Artificial Intelligence (AI), drawing on lessons from major research programmes, industry practice, and cross-national collaboration. It emerges from a partnership between Responsible AI UK (RAi UK) and the European Trustworthy AI Association, reflecting a shared commitment to advancing responsible, inclusive, and effective AI governance.

The report synthesises insights from recent projects, including Confiance.AI and the UKRI Trustworthy Autonomous Systems (TAS) Programme, alongside discussions from an April 2025 international workshop. It explores four core themes: operationalising trustworthy AI in practice; building inclusive AI through co-design; trust and accountability in AI systems; and governance and regulation.

Key findings highlight a broad consensus on foundational trustworthy AI principles—transparency, fairness, accountability, explainability, privacy, human oversight, and robustness—yet persistent challenges remain in translating these into operational practice, especially in high-stakes and regulated sectors. Current toolkits and frameworks are valuable but often lack specificity for sectoral compliance, require costly and complex audits, and are unevenly adopted, particularly among small and medium enterprises.

Emerging challenges include the unpredictability of large language models (LLMs), the difficulty of certifying general-purpose AI, divergent international regulatory regimes, deepfake proliferation, and the risk of cultural homogenisation. Opportunities lie in scaling collaborative approaches, developing adaptive and risk-proportionate governance frameworks, enhancing interdisciplinary skills, and embedding meaningful user participation throughout AI's lifecycle. The paper stresses that trust is contextual, dynamic, and distinct from trustworthiness, requiring calibrated approaches to both technical robustness and societal expectations.

By consolidating practical experience, policy insight, and research evidence, this report aims to inform actionable strategies for regulators, developers, and civil society. It calls for sustained international cooperation, sector-specific frameworks, lifecycle monitoring, and public engagement to ensure that AI systems serve diverse communities equitably, safely, and effectively.

Introduction

Responsible Ai UK and the European Trustworthy AI Association have formed an official partnership. To mark this partnership and to advance collaboration, we hosted a workshop on 8 April 2025 bringing together researchers, policymakers, and industry leaders from Europe, Canada, and the UK to share insights from national research programmes and develop a common understanding of emerging challenges in developing Trustworthy AI.

We aim to foster international collaboration, identify challenges, and produce actionable recommendations for researchers, industry leaders, and policymakers. The workshop also highlighted lessons learned from key research programmes such as Confiance.AI, the UKRI Trustworthy Autonomous Systems Programme (TAS), and other relevant initiatives.

Substantial work has been done by the partnership and related organisations. We aim to bring together findings from that work, to put learning into industry practice, accelerate adoption and support businesses and other organisations in achieving compliance with regulation and adoption of best practices.

Collage of photos from the joint RAi UK and European Trustworthy AI Association event, *European AI Responsibility: Policy, Innovation, and Ethics Across Borders,* Royal Academy of Engineering, London, 8 April 2025

Context

Lessons learned from projects on trustworthy AI

In recent years, there have been a number of projects that have looked to build responsible AI. Toolkits and frameworks have been produced to support the design, development, and governance of AI to ensure it is trustworthy by design, and that there is confidence in that trustworthiness by government, industry, and the public.[1] In advance of the joint workshop we reviewed these to identify lessons learned and emerging challenges and opportunities. These results are based on a need for operationalising the development of trustworthy AI-based safety critical systems.

The prospects for trustworthy AI are positive. There is emerging consensus across organisations and jurisdictions on fundamental trustworthy AI principles: transparency, fairness, accountability, explainability, privacy, human oversight, and robustness. The EU Ethics Guidelines, OECD framework, and UK approach all emphasise similar core values, suggesting a mature understanding of what trustworthy AI should embody. However, while ethical principles are well-established, translating them into practical day-to-day operations remains challenging. This gap is particularly pronounced in regulated sectors like healthcare and law, where it can be unclear what responsible design and deployment actually look like in practice (especially in life critical systems); and where the translation of regulation into Computer Science language and requirements invite ambiguity. It is evident that further work at the interface of these two disciplines is required in order to ensure that regulatory ambition and the technically possible combine to deliver policy intent effectively.

Trust in AI encompasses multiple dimensions including fairness, accountability, transparency, and explainability—not just accuracy. However, measuring these dimensions remains difficult, concepts are nebulous and open to differing degrees of interpretation and implementation, and there are often trade-offs (for example, improving transparency can sometimes reduce model performance). Involving domain experts, regulators, and end-users from the start of AI development significantly improves system trustworthiness and usability. However, meaningful co-production remains the exception rather than the norm, especially for foundational models, general purpose AI systems (GPAI), and large language models (LLMS) being deployed without adequate input from actual users. A lack of user-centric design and input compounds challenges of trustworthy and responsible development and deployment by potentially overlooking underrepresented or marginalised groups; potential skill or process gaps between the end-user and the application; and where interacting with consumers, failing to understand and appreciate impacts of trade offs on nebulous duties (such as the UK FCA's consumer duty); limiting the impact of co-production.

Building trustworthy systems demands combining legal, social, technical, and policy expertise, combined with an understanding of the impact when interacting with humans; but this collaboration needs deliberate facilitation and governance structures to be effective. Even at the level of a single organisation, these considerations need to be translated into organisational transformation to facilitate interdisciplinary cooperation at a technical level in order to have effective impact. Early commitment to ethical frameworks is more likely to deliver better results than reactive responses, and these in turn should reflect the experience of real-world use cases to ensure operationalisation can take place .

While general-purpose toolkits for bias detection, fairness assessment, and explainability are becoming available, they often don't meet the specific certification and compliance requirements of high-stakes sectors or deliver a degree of robustness and specificity to deal with demanding legislation on the scale of the EU AI Act and the accompanying technical standards which will follow. Specialised tools aligned with regulatory standards and industry practices are still needed, as is a framework to ensure that these tools meet minimum quality requirements, and interoperability standards are provided for.

Maintaining trust throughout an AI system's lifecycle requires ongoing updates, feedback loops, and regular monitoring. Many systems are deployed without clear plans for continuous evaluation, which is problematic as data, regulations, and practices constantly evolve. In addition, a lack of systematic operational and functional specifications makes verification and validation more problematic, and in turn unable to provide results that can be trustworthy. Responsible AI deployment requires structured governance that matches oversight intensity to risk level, combining technical safeguards with human oversight and continuous adaptation mechanisms.

Emerging challenges and opportunities

1. LLMs are at the forefront of AI adoption in mainstream organisations. There is now better collective understanding of the limitations of LLMs, which should help to address the challenges they present to trustworthiness, but mass uptake also means that issues can arise in many contexts. LLMs create **many challenges to trustworthiness.** They generate factually inaccurate statements. Their probabilistic nature limits their ability to perform contextual, discretionary decision-making that requires genuine analytical thinking. They can struggle to know when not to answer, or flag ambiguous outputs, especially with multimodal and open-ended tasks. They lack genuine understanding or intent, but can give users the impression that they are capable of doing so. Users may develop inaccurate expectations from systems that sound highly human but fail in non-human ways. It may be necessary to redefine or expand concepts of trustworthy AI to include qualities like calibration of confidence, deception avoidance, or continuity of AI behaviour over time .

DAISY – Diagnostic **AI Sy**stem for Robot-Assisted A&E Triage, was a TAS funded project.
The collaboration between the Universities of York and Southampton and the York and Scarborough Teaching Hospitals worked on a prototype robot-assisted A&E triage solution for reducing patient waiting time and doctor workload - https://tas.ac.uk/research-projects-2022-23/daisy/

Prof Sarvapali (Gopal) Ramchurn, RAi UK CEO, Karla Quintero, ETAIF and others at the Trustworthy AI Summit in Paris, 16 September 2025.

2. There has been significant progress in the safeguarding and assessment of classical Machine Learning (ML) systems based on **the notion of operational design domain** – which defines the specific conditions for an automated system to operate safely and effectively. However, there remain gaps in assessment and verification. It is far more difficult to assess the trustworthiness of large, adaptable general purpose AI models (like LLMs) for their vast, unpredictable range of potential uses, which challenge traditional domain-specific certification. Traditional software testing principles fail for complex neural networks, necessitating entirely new verification and validation methods. Current trustworthiness evaluations require extensive manual effort, creating bottlenecks for the volume of AI systems needing evaluation. Moving successful interdisciplinary approaches from specific projects or hubs (like TAS, Confiance.ai, and ZERTIFIZIERTE KI) into mainstream, large-scale industry practice remains a significant challenge.

3. There are also challenges in governance and regulation. Differing AI policies across countries (EU AI Act, US Blueprint and AI Action Plan, PRC Guidelines) create complex compliance challenges for global organisations. The lack of universally accepted definitions and attributes of trustworthiness hampers consistent evaluation, though positive steps are being taken in this direction with forthcoming ISO standard (22989) and the EU AI Act and associated ETSI standards. Clear mechanisms for AI assessment organisation approval and compliance monitoring are still missing, as is a common international legal language, and integrated legislative, regulatory and technical landscape.

4. Tools that enable sophisticated deepfake and synthetic identity scams are outpacing traditional defence mechanisms. There is a risk of creative homogenisation: over-reliance on generative AI threatens cultural nuance. Intellectual property questions and in particular AI training on copyright materials remain unresolved.

5. Implementing trustworthy AI encounters barriers. **Compliance audits are costly and require expertise and access to resources that smaller organisations often lack.** Many AI professionals view governance as an unwelcome burden. There are two solutions required here – one is to ensure that regulation and required governance is reasonable and proportionate to both the use case and the capacity of the developing or deploying organisation; and that the value of meaningful governance and regulation needs to be changed through training and education. Technology and social expectations evolve quickly, making governance frameworks difficult to maintain and adapt over time.

Collaboration offers ways forward. There is growing recognition that combining legal, social, technical, and policy expertise strengthens AI governance. Tools like "Framework and Self Assessment Workbook for Including Public Voices in AI"[2] enable broader stakeholder involvement in development processes. Learning from past successful collaborations can contribute to framework interoperability.

There are opportunities to reduce manual effort in trustworthiness evaluations through automated tools and processes. Simplified automated tools could make fairness and bias compliance testing more accessible to smaller organisations. There is a need to make frameworks adaptive, to evolve with technological, regulatory and societal changes, and for ways to better fit measures to risks.

6. There are different opportunities in different sectors. There is increasing recognition that government AI applications require specialised frameworks different from private sector approaches. Education and training programs can help AI developers better understand ethical implications and help companies see the benefits of creating a responsibly focussed approach to AI. More can be done to translate research advances into practical implementation guidance.

7. We need better understanding of barriers to industry adoption of these toolkits, of what measures are successful in incentivising their use, and of where current tools fall short in practice and what improvements could make them more usable for product teams. What organisational structures or workflows truly merge human-centred design with technical innovation in AI? How can accountability gaps be avoided when multiple parties and AI subsystems are involved in an outcome?[3]

8. Trustworthy AI requires collaboration across research disciplines and across borders. What are the most promising models for collaboration we should strengthen or replicate (international research consortia, public-private partnerships, regulatory sandboxes)?

Consider the example of the TAS Hub – how might we create a sustained, international equivalent that continuously drives trustworthy AI research and policy advice? How can we better link research efforts with policymaking so that scientific insights rapidly inform governance, and vice versa, in this fast-moving field? One of the initiatives moving in this direction is the European Trustworthy AI Association. This association has begun to take on the challenge of fostering international cooperation across different sectors and disciplines. It operates based on the principles of European governance. The main goal of the association is to pool and align the efforts of research, standardisation, and industrial ecosystems. By doing so, it aims to ensure that future developments of AI-based systems are trustworthy by design. The association seeks to enable fundamental research to respond to the needs of industry and apply its findings to real-world use cases; in this sense, it also facilitates the creation of collaborative projects addressing these needs. All these activities are to be carried out in coherence with standards and should further propose precise content for future standardisation initiatives. This is done as a collective effort, reaching preliminary consensus among the involved parties.

1. Examples can be found at https://catalog.confiance.ai
2. Patel, R. (2025) A Framework and Self Assessment Workbook for Including Public Voices in AI. Elgon Social Research and ESRC Digital Good Network. Available at: https://elgon.social/framework-and-workbook > Accessed 31/07/2025
3. RAi UK (2025) Frameworks and Toolkits for Assuring Responsible AI, White Paper 02 August 2025

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

Theme 1 - Operationalising trustworthy AI in practice

Operationalising trustworthy AI requires context-specific applications while maintaining systematic approaches to evaluation and learning. Success depends on bridging technical optimisation with ethical considerations, ensuring meaningful public engagement, and creating robust mechanisms for learning from both successes and failures. The challenge lies in balancing the need for technical specificity with the flexibility required for diverse applications and cultural contexts.

The translation of abstract AI ethics principles into practical, actionable frameworks remains complex. Generic frameworks often prove inadequate for specific use cases, with participants noting that available methodologies only address portions of complex problems. There are critical gaps between theoretical principles (which can be expressed as text and encompass logic, constraints, ontologies, ensembles etc.) and end-to-end implementation (the real world made of continuous data flows). Grounding the former in the latter and linking principles and measurable factors remains a human task to be done on a case-by-case basis.

The challenge extends beyond technical implementation to ensuring that compliance assessments are meaningful rather than superficial "box-ticking" exercises.

AI failures present unique challenges compared to traditional software bugs, as they often have broader societal impacts that cannot be addressed through conventional engineering approaches. The field lacks robust post-failure processes for systematically capturing, tracking, and learning from failures. There is a recognised need to balance prevention strategies (through ethical regulations) with protection mechanisms for those harmed by AI systems, while developing transparent processes for sharing lessons learned with society.

Current informatics education largely omits AI ethics and data management, with computer scientists typically trained to optimise systems without considering ethical dimensions. This creates a cultural mindset that disregards consequences. Meaningful interdisciplinary collaboration requires developing common vocabularies, fostering expert willingness to work outside comfort zones, and building mutual understanding between disciplines.

Existing regulations often struggle to balance addressing outcomes and AI processes. More often than not, this balance favours the former over the latter, creating flexibility but also uncertainty. The favouring of this flexibility of interpretation places reliance on litigation-based clarification, leaving critical decisions in judges' hands, with limited meaningful technical insight and input. Small and Medium Enterprises particularly face compliance challenges due to associated costs, while the technical reality that machine learning models can never achieve 100% accuracy complicates regulatory frameworks.

Trust in context

Trust in AI systems is contextual, situational, and dynamic rather than binary or permanent. A system might be trusted in one domain but not another, depending on how well its limitations are understood. Trustworthiness has boundaries tied to operational design domains and contains assumptions based on the distribution of training data. General-purpose models face particular challenges here, as they are deployed beyond their original operational boundaries.

Trustworthy AI must be addressed at the system level, recognising that AI systems comprise multiple components created by different organisations. This distributed development inevitably complicates trustworthiness, transparency, and alignment throughout the AI value chain, and generates issues around liability and responsibility for the facilitation of compliance. The challenge is compounded by the iterative and evolutionary nature of systems development, requiring trustworthy mechanisms to evaluate how well systems meet their intended purposes as they evolve over time.

There are gaps between technical development and cultural readiness. People express concerns about surveillance and misuse. Public trust may be eroding further over time, driven by feelings of helplessness. Systems may then struggle to gain public acceptance even before implementation.

The burden of ethical AI implementation is unevenly distributed across the industry. Large tech companies often have formal ethics committees, while smaller developers lack the same infrastructure despite being expected to adhere to the same responsible innovation principles. Ethical preferences are a spectrum, and individual organisations' ethical preferences, red-lines, and balance of risk will inevitably vary significantly. Globally, regulatory approaches vary significantly, with UK/EU emphasising strict regulation while Asia and the Middle East show greater leniency, and the United States is pursuing an innovation and growth first approach.

Current AI evaluation methods face significant limitations, although there are positive attempts to improve evaluation, such as the Stanford AI index[4] which annually presents new benchmarks, while informal leaderboards such as Hugging Face[5] assume the role of a globally available standard. Benchmark comparisons between leading tech companies undermine the validity of performance claims. The field possesses multiple metrics but lacks a coherent structure linking metrics to regulatory compliance and standards, or a formally established standard for evaluating certain types of AI outputs (such as audio similarity), making accountability difficult to establish. Risk classification frameworks may overlook harms in apparently "low-risk" platforms, such as social media, where indirect consequences such as disinformation and mental health impacts may nevertheless come into play.

Ways forward

The IDEAL framework for surgical robotics[6] offers a structured model for technology validation through staged implementation: preclinical testing, early human trials with iterative learning, and larger-scale randomised trials with close monitoring. This provides a clear, step by step process for building trust, monitoring risks, and embedding accountability. Shared standardised learning frameworks that capture everything that happens, both successes and failures, for common future benefit. This approach shows how systematic documentation and learning can improve outcomes while maintaining safety.

New tools are constantly being developed and iterated to help translate key terms across sectors, addressing the lack of shared terminology. Creating common vocabularies and fostering environments where experts can collaborate outside their comfort zones can be significant steps toward implementing trustworthy AI practices.

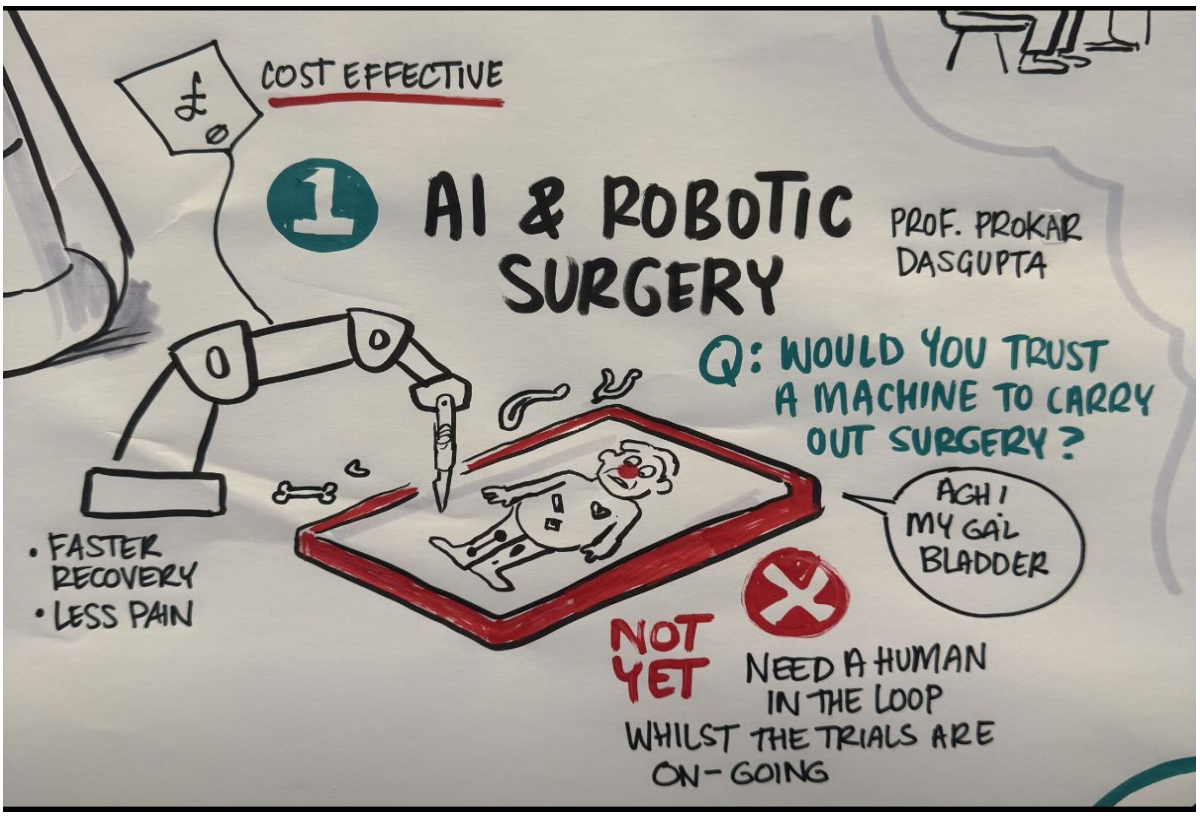
Other frameworks, such as the "End-to-End Methodology for Engineering Trusted ML-based Systems" have been produced on the industrial scope in the Confiance.ai project[7] and continue to evolve within the European Trustworthy AI Association. This framework addresses the multi-sector/multi-disciplinary challenges, and many methodological guidelines derive from it. The body of knowledge (https://bok.confiance.ai/) allows navigating the method through a cycle that emerges from revisiting the classical engineering disciplines (Systems Engineering, Software Engineering) in order to be able to handle correctly, and through the trustworthiness lens, the integration of the ML technology in critical systems. Fraunhofer IAIS have also produced an assessment catalogue to assess (and safeguard) AI systems, with an emphasis on quality and trust as competitive advantages in the responsible assessment of AI systems, supported by structured guidance to define application specific assessment criteria.[8]

One key takeaway is that commonalities and complementarity among frameworks such as IDEAL, the End-to-End Methodology for Engineering Trusted ML-based Systems, the Fraunhofer IAIS assessment catalog, and other initiatives (even if they only entail partial contributions to the entire development cycle) could be studied for mutual enrichment.

Other methodologies were highlighted which demonstrate promise and value-add within the ecosystem: Google's Data Cards Playbook;[9] and Data Feminism's framework for assessing and addressing power imbalances in technology;[10] IBM's Responsible AI tools;[11], ZERTIfizierte KI (Certified AI) assessment and certification platform;[12] and other bottom-up meets top-down methodologies have all shown promise.

The key is developing generic frameworks that can be applied to specific contexts while maintaining domain-specific understanding.

4. https://hai.stanford.edu/ai-index
5. https://huggingface.co/
6. https://www.nature.com/articles/s41591-023-02732-7; https://www.ideal-collaboration.net/news/the-ideal-robotics-colloquium-a-new-framework-for-the-development-evaluation-and-long-term-monitoring-of-surgical-robotics/
7. https://catalog.confiance.ai/records/n6ag2-b8q77
8. https://www.iais.fraunhofer.de/en/publications/studies/2023/ai-assessment-catalog.html
9. https://sites.research.google/datacardsplaybook/
10. https://data-feminism.mitpress.mit.edu/
11. https://www.ibm.com/solutions/ai-governance#:~:text=A%20toolkit%20that%20seamlessly%20integrates,Google%20Vertex%20and%20Microsoft%20Azure.
12. https://www.zertifizierte-ki.de/



Despite a fully autonomous robot that is 100% accurate in removing pig galbladders, recently reported, the public say "Not yet". Credit: Professor Prokar Dasgupta

Theme 2 - Building inclusive AI through co-design

AI systems are increasingly used to make decisions that significantly impact people's lives, including in employment and justice, but can lack transparency and be developed without meaningful involvement of users. Minority groups and already marginalised populations are facing compound exclusion through intersectionality. LLMs can foreground the views of some groups at the expense of representing others.

The underlying challenge is that the world is a diverse place. Decision-making AI models' and LLM performance is highly context-sensitive, varying significantly based on organisational approaches, regional differences, and cultural contexts. One-size-fits-all approaches systematically fail to account for this diversity, leading to systems that work well for some populations while failing others.

Specific populations face particular challenges in AI systems. Refugees, for example, with different exposures to AI technology, children who may not understand the difference between AI and human interaction, and marginalised communities whose data and perspectives are often absent from training datasets, may face additional needs when interacting with systems designed for a majority population use case. Datasets themselves may also reflect population challenges: medical data, for example, over-represents US populations, while minority groups, non-verbal cues, and cultural contexts are frequently overlooked.

Despite the recognised need for interdisciplinary collaboration to improve co-design methods, there can be limited appetite from experts to cross disciplinary boundaries. Professionals lack training in interdisciplinary work, and there is insufficient integration of perspectives from social sciences, arts, and humanities. This creates silos that prevent holistic understanding of user needs and impacts.

Industry shows reluctance to invest in user involvement at the design stage due to funding constraints and lack of incentive structures for inclusive design approaches. The competitive race framing of AI development can trap organisations into losing their unique value propositions, as seen in the music industry. Companies often prioritise speed and cost over meaningful user engagement.

Market research approaches to understanding user needs carry significant risks, as users may not fully understand the risks involved or recognise potentially harmful features. The Facebook "like" feature's psychological impact on teenagers exemplifies how seemingly benign features can cause harm. Traditional user feedback methods often fail to capture the complexity of AI's social impacts.

Computer science professionals can be uncomfortable with open scrutiny and lack training in ethical considerations and human-centred design. There is a critical need to shift from purely technical coding to meaningful interaction with people. Ethics education, while recognised as important, remains broadly defined and poorly integrated into technical training.

New gaps and risks are emerging. A significant generation gap exists in AI understanding, with children using generative AI while parents lack comprehension of these tools. This creates crossover challenges where no one takes responsibility for ensuring appropriate, age-appropriate design. All progressive universities should integrate AI literacy into their curricula to prepare future professionals effectively.

There are growing concerns about over-reliance on AI systems leading to cognitive decline, particularly among new generations of professionals who may lack critical thinking skills compared to older cohorts. This dependency affects professional judgment and decision-making capabilities across sectors.

The goal is not just technical improvement but both organisational and social transformation—creating AI systems that serve diverse communities equitably while maintaining safety, trustworthiness, and effectiveness. This requires sustained institutional support, interdisciplinary collaboration, and a fundamental shift from technology-centred to human-centred design approaches.

Ways forward

Healthcare demonstrates successful integration of stakeholders through the "3Cs" model (companies, civil society, and countries). Public and Patient Engagement (PPE) is now a standard requirement in healthcare grant applications, with increasing numbers of patients sitting on grant evaluation panels. The UK National Institute for Clinical Excellence (NICE) serves as a positive example of balancing cost-effectiveness, performance, and end-user involvement in evaluation.

Co-production with domain experts leads to more trustworthy and usable AI systems, as demonstrated in TAS Hub and healthcare AI projects. When training models, especially large language models, using small, relevant datasets can enhance performance and relevance for specific communities rather than relying on massive, generic datasets.

Rather than pursuing AI models that are "everything to everyone," focused-purpose AI systems may be more useful and accurate. A federation of interacting models might better include and contain conflicts while respecting different cultural contexts and needs. This approach recognises that "responsible" and "trustworthy" may mean different things in different contexts.

Open-source solutions can enable community adaptation and customisation, as can engaging users early in development processes to ensure usability and relevance, applying tailored questioning techniques when consulting users from different backgrounds. Allowing for regional customisation can avoid risks of relying on globally uniform models.

Restructuring research grants can incentivise interdisciplinary work packages. Educational reforms can promote training in interdisciplinary methods, and integrate perspectives from social sciences, arts, and humanities more systematically.

We need better evidence of what works in participatory methods that meaningfully and specifically include excluded groups. More evaluation of how co-design impacts AI trustworthiness could provide lessons.

Data disclosure and privacy remain major concerns, with users often lacking visibility into what is collected and how it is used. Regulations are crucial: trust is not just a product of good engineering but of transparent evaluation and oversight processes. Companies should be held accountable for "RAI washing" through common signs and characteristics.

New governance models require legal incentives and penalties that reward companies for transparency and socially positive outcomes. Actions supporting responsible AI goals should be protective and net positive for companies, moving beyond current concerns about liability and fraud to focus on public service.

While there is much to be optimistic about the awareness of these risks and potential solutions, global consensus on both the importance of these issues, and the possible avenues to resolving them remains elusive. The United States Government's AI Action Plan recommends taking proactive steps to remove references to diversity, equity and inclusion (DEI), identifying this as introducing social engineering into systems. Those considering policy responses, and those considering adoption, will need to weigh on balance the impact of adhering to and pursuing a proactive approach to co-production, with very diverse approaches and responses to the challenge in a transatlantic context.

Theme 3: Trust and accountability in AI systems

Trust in AI systems defies simple definition and measurement because it is inherently nebulous. It means different things to different people in different contexts. It can be affected by many factors including the reputation of the system or company, previous experiences of individuals and groups, and the specific domain of application. What constitutes trustworthy behaviour in one sector or use case may not apply in another, making standardised trust measures challenging to develop and implement; for example, engagement is an oft-cited metric for demonstrating trustworthiness, but fails to reflect deeper concerns regarding fairness and accountability. This contextual dependence extends to how trust is built over time, with systems having histories that influence current perceptions and expectations.

The conceptual understanding of trust has evolved beyond simple accuracy measures to encompass Fairness, Accountability, Transparency, and Explainability (FATE) alongside privacy, human oversight, and robustness. This reflects growing recognition that trust in AI systems requires multiple dimensions of assessment, including how well systems perform their intended functions, their openness about capabilities and limitations, and their alignment with user expectations and societal values.

People may apply different standards to human and machine decision-making, judging humans by their intentions and impacts, but machines solely by the impact of their actions. They may also bring heightened expectations for consistency, transparency, and accountability of AI systems that may not apply to human decision-makers in similar contexts.

The dynamic nature of trust means it can be built, eroded, or destroyed based on ongoing performance and changing circumstances. Trust may be moderated by existing confidence in systems before AI integration, suggesting that AI trust-building must consider the broader socio-technical context rather than focusing solely on the AI components.

Trust is different from Trustworthiness, which is defined by ISO as the ability to meet stakeholder expectations in a demonstrable, verifiable, and measurable way. Trust involves both logical reasoning and emotional components, making it inherently subjective and difficult to quantify directly. Measuring trust presents significant methodological challenges because it requires asking people about their perceptions and understanding what their responses mean across different stakeholder groups. Rather than measuring trust itself, practitioners often rely on measuring derivatives of trust such as user engagement, customer complaints, evidence-based performance, and the effectiveness of and confidence in verification systems like auditing and certification processes.

The 26 forbidden discrimination criteria in France

Compliance against technical standards, often verified by certifications, helps to build consumer confidence, though their effectiveness varies across contexts and applications. The challenge lies in developing measures that capture both quantitative performance and qualitative user experiences while accounting for the full range of stakeholders affected by AI systems.

Fairness can sometimes be comparatively easier to measure when it means equal treatment for different people and groups. Perception of fairness may offer more concrete assessment pathways than trust itself.

Building trust involves aligning expectations with system capabilities from the design phase onward, recognising that clients do not always know what they want from AI systems.

Clear definition of system requirements enables proper measurement and helps establish realistic performance expectations. Understanding the trustworthiness of alternatives provides necessary context for comparison, particularly since AI systems typically replace human activities that operate under different judgment standards.

Communication of uncertainty to the public requires context-appropriate approaches, with trust building extending beyond the AI system itself to include confidence in the processes and people who verify these systems. How uncertainty and limitations are communicated directly affects how systems are used and perceived. This communication challenge becomes particularly acute when dealing with complex technical concepts that must be made accessible to diverse audiences with varying levels of technical expertise.

The assumption that trust is inherently good and that more trust is better, requires careful examination. In some contexts, less trust or more critical trust can be appropriate. Calibrated trust involves understanding when to trust systems and when to maintain scepticism, requiring users to develop AI literacy alongside traditional media literacy skills. This approach recognises that trust should be situational and based on understanding rather than blind acceptance. Building appropriate trust requires helping users understand the basics of how AI systems are trained, their inherent biases, opportunities, and limitations.

.

One of the leads to address this challenge could be similar to the approach of the AI Trust Alliance. The alliance includes several actors, including the Institute for Electrical and Electronics Engineers (IEEE), and pushes for a consensus on an AI Trust Label based on a specification built on top of the 7 pillars of the EU AI Act's approach on trustworthy AI: Human Agency & Oversight, Technical Robustness & Safety, Privacy & Data Governance, Transparency, Diversity, Non-Discrimination & Fairness, Societal & Environmental Well-Being, and Accountability. Through this collaboration, an AI Trust Label would rely on more than subjective acceptance, providing a high-level ranking that should then give end-users the possibility for further understanding on the criteria and metrics that led to the ranking.[13]

Accountability

Supply chain complexity and dispersed ownership for different parts of systems complicates liability, transparency, and trust assessment. When combining elements including datasets and libraries, the result can be an entangled picture where it is not simple to know everything that could affect performance and outcomes or to allocate liability and exercise responsibility appropriately.

Recent research on explainability of LLMs reveals that explanations given do not necessarily clarify how or why models behave as they do, which compounds the problem of assessing how much they can be relied on. LLMs can also produce convincing but factually incorrect outputs, leading to "overtrust" by users who may not have the expertise to critically evaluate AI-generated content. They may also alter their behaviour when they detect evaluation, complicating causal analysis, fairness assessment, and optimisation efforts. This behavioural complexity requires organisations to implement mechanisms for questioning explanations and avoiding blind trust while encouraging appropriate scepticism as part of responsible use.

The need for monitoring AI systems throughout their lifecycle parallels approaches used in drug development, where monitoring continues through development stages and after deployment. Unlike drugs, however, AI models may continue to change after deployment, creating ongoing drift and unknown effects. AI development must embed safety proactively and iteratively, rather than relying on retrospective fixes, and address their ability to change and drift over time. In defence and security, online learning and adaptation over time during operation is strictly forbidden in order to ensure maximum accountability.

Different trust measures may be relevant at different stages of the AI lifecycle, from development through deployment to ongoing operation.

The lack of standardised monitoring approaches for AI systems, unlike established practices in pharmaceutical development, represents a significant gap in current trust-building mechanisms. Addressing this gap requires developing systematic approaches to lifecycle monitoring that address the unique characteristics of AI systems while building on established practices from other domains.

Public service and regulatory contexts

In public service contexts, users tend to focus on the public body rather than technology providers when assessing trustworthiness, expecting public organisations to take on all liability and ensure system trustworthiness. This expectation creates unique challenges for public sector AI deployment, where complex procurement and implementation processes may obscure accountability lines. The development of "information sheets" similar to medical contraindication warnings could help provide employees and users with necessary information about potential secondary effects and risks.

The public sector could lead in practice as well as in policy. The UK Government is relatively active in supporting responsible AI, but it should share learning more effectively. Recently it was reported that 11 of 12 AI trials in the public sector have been stopped or abandoned. Lessons should be shared through mandatory publication of results, including challenges, failures, and reasons, providing valuable learning for the broader community.

13. https://www.trustalliance.ai/.

Theme 4: Governance and regulation of AI

Understanding how to apply existing and new regulations to novel applications of AI is not simple, and we have relatively little history to work with. It is often not entirely clear what users of specific AI applications need to do to comply fully with the requirements of current regulatory frameworks, including the EU AI Act (although technical standards for the Act will be introduced by ETSI in time for the Act's full entry into force in August 2026). This adds another layer of uncertainty to trust-building efforts. Additionally, the assumption that standards will apply uniformly may not match AI system capabilities, creating gaps between regulatory expectations and technical realities that must be addressed through careful implementation and ongoing assessment.

Clear rules can set expectations and provide legal certainty that potentially boosts investment, but burdensome regulations create barriers to entry, particularly for small and medium enterprises. Conversely, the absence of regulation or enforcement can disadvantage responsible AI developers who voluntarily adopt ethical practices. Governance frameworks should be designed to capture domain-specific context over time and adapt to evolving threats and opportunities. This is a complex time horizon that acknowledges both the rapid pace of technological change and the slower evolution of underlying social and ethical challenges. Ideally, public policy and regulation would be responsive to new developments while collecting learnings from diverse contexts and generalising insights across sectors and countries.

The rapid pace of innovation and uptake raises questions about how accurately future challenges can be anticipated and legislated for. Typically, technological progress occurs in sprints of 6 to 12 months, while the lead-in time for regulatory change or technical standards development stands at 3 – 5 years. Policy-makers are trying to form effective measures to apply existing principles under substantially new conditions, which may not provide developers with the certainty they would like. Even regulated industries like water, energy, and aviation still experience regulatory failures, highlighting the need for continuous adaptation. Governance should serve public well-being rather than political or corporate interests.

AI governance can be made more effective by learning from AI failures, but that depends on the detail of failures being made public, which does not always happen.

The European Union's General Data Protection Regulation (GDPR) has been criticised as too blunt an instrument, while the EU AI Act faces criticism for focusing solely on risks without adequately considering benefits. Risks cannot be reduced to zero, and society's risk tolerance levels must be informed by careful analysis of benefits and trade-offs.

A significant asymmetry of power exists between major technology companies and regulators, arguably representing a market failure that demands regulatory intervention. Co-regulation approaches could help overcome information asymmetries between industry and government, as demonstrated by Singapore's development of regulatory mechanisms in conjunction with the private sector. Testing regulatory models in controlled environments like regulatory sandboxes helps determine whether frameworks are fit for purpose.

Regulations often may not align with a system's technical capabilities, creating gaps between legislative requirements and current technical ability to comply or report. Ideally, this should prompt an engagement process to make compliance achievable and improve mutual interpretation of the law, though companies are often inclined to avoid liability rather than to engage constructively.

Different sectors face varying challenges in AI regulation and compliance. In medicine, there are established processes for identifying when procedures go wrong, but in areas like credit scoring, unfair outcomes may only become apparent when identifiable groups are systematically disadvantaged over time. This can be socially sanctioned: as the public, we tend to tolerate a degree of luck, some variation in outcomes, as long as no group is repeatedly and systematically disadvantaged. This delayed detection means many people may be harmed before problems are noticed.

The insurance sector might provide an interesting lens for understanding AI governance. Insurers need to price risks in corporate applications of AI. It would be valuable to know more about how they do that, and which risks they can price, and which tend to confound their risk assessment models.

New challenges are emerging around AI agents, hostile cybersecurity bots, and potential impacts on financial systems. The question of AI-assisted versus AI-generated content raises important questions about acceptability and liability. The systematic appropriation of copyright works and the moral implications of removing humans from creative processes remain contentious. The reusability of AI models present additional regulatory challenges, particularly around security aspects and dynamic risks that require models to keep evolving.

Progress is being made in regulatory standardisation through frameworks like the EU AI Act, NIST AI Risk Management Framework, and the UK's five core AI principles. However, fragmented legal and policy landscapes across regions make global compliance difficult.

Governance and regulation can only do so much and need to be in step with society. Preparing society for AI's continued evolution requires public education about AI systems and the creation of the necessary governmental infrastructure to do so. There is an international consensus on this approach, through the creation of both the global network of AI Safety and Security Institutes, and the series of AI Summits and their thematic focus on safety, action, and impact.

We are all still too systematically ignorant about AI governance, failing to transfer lessons across contexts, sectors, and countries as well and as fast as we might. Partnerships for sharing challenges and solutions like this one are valuable but given the number of countries and sectors currently working out how to use AI well, sharing perspectives and experiences could be done much more to help us all catch up with developments in AI capability. One area we could share ideas on more is how to prepare and inform society about what is happening now and what may be coming next .

Conclusion

Designed by Freepik.

Trustworthy AI will not be achieved through principles alone, nor through regulation in isolation, but through a deliberate, collaborative effort to embed ethical, technical, and societal considerations into the full lifecycle of AI systems. The path forward demands that we close the gap between aspiration and implementation, share lessons openly across borders and sectors, and equip both industry and society to navigate the rapid evolution of AI capabilities. Partnerships such as that between RAi UK and European Trustworthy AI Association demonstrate the value of aligning research, policy, and practice to create frameworks that are adaptable, context-sensitive, and grounded in real-world use. By acting collectively and proactively, we can shape AI systems that are not only innovative, but also genuinely worthy of the trust placed in them.

Further reading

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

Trustworthy Autonomous Systems (TAS) Hub
https://tas.ac.uk/

Confiance.ai
https://www.confiance.ai/
https://www.confiance.ai/foundation/
HAL Collection: https://hal.science/CONFIANCEAI

European Trustworthy AI Association
https://www.trustworthy-ai-association.eu/

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
https://www.fraunhofer.de/en/research/fraunhofer-strategic-research-fields/artificial-intelligence.html

ZERTIFIZIERTE KI
Fraunhofer IAIS is working with the German Federal Office for Information Security (BSI), the German Institute for Standardization (DIN), and other research partners to develop assessment procedures for the certification of artificial intelligence (AI) systems. The aim is to ensure technical reliability and responsible use of the technology. Industrial requirements are taken into account through the active involvement of numerous associated companies and organisations representing various sectors such as telecommunications, banking, insurance, chemicals, and retail.

The project activities include the development of AI assessment criteria and AI assessment tools, as well as the transfer of results into standardisation. In addition, it investigates new business models and markets for AI testing and AI certification. The project also takes a holistic approach, incorporating legal and philosophical-ethical issues.

Regulation as an incentive for behavioural change (i.e., need for clear requirements for compliance)

- Black, J., & Murray, A. D. (2019). Regulating AI and machine learning: setting the regulatory agenda. European journal of law and technology, 10(3).

Reputational challenge (i.e., when the firm's reputation is challenged because of bad publicity)

- Prakash, Aseem (2000), Greening the Firm: The Politics of Corporate Environmentalism, Cambridge: Cambridge University Press
- Gunningham, Neil, Kagan, Robert A. and Thornton, Dorothy (2003), Shades of Green: Business, Regulation and the Environment, Palo Alto, CA: Stanford University Press.
- Mehta, Alex and Hawkins, Keith (1998), 'Integrated Pollution Control and Its Impact: Perspectives from Industry', Journal of Environmental Law, 10, pp. 61–77.
- Kagan, Robert A., Gunningham, Neil and Thornton, Dorothy (2003), 'Explaining Corporate Environmental Performance: How Does Regulation Matter?', Law and Society Review, 37, pp. 51–90.

Reluctance to commit resources to toolkits if future regulations demand different parameters

- [Regulatory uncertainty as a compliance cost]: Cordes, J. J., Dudley, S. E., & Washington, L. Q. (2022). Regulatory compliance burdens. https://regulatorystudies.columbian.gwu.edu/sites/g/files/zaxdzs4751/files/2022-10/regulatory_compliance_burdens_litreview_synthesis_finalweb.pdf

Lack of Enforcement and Pressure to Comply

- Thornton, Dorothy, Kagan, Robert A. and Gunningham, Neil (2005), 'General Deterrence and Corporate Environmental Behavior', Law and Policy, 27, pp. 262–88.
- Gunningham, Neil, Thornton, Dorothy and Kagan, Robert A. (2005), 'Motivating Management: Corporate Compliance in Environmental Protection', Law and Policy, 27(2), pp. 89–316
- Mendeloff, John and Gray, Wayne (2004), 'Inside OSHA's Black Box: What is the Link Between Inspections, Citations and Reductions in Different Injury Types?', Law and Policy, 27, pp. 219–37.
- Shimshack, Jay and Ward, Michael (2005), 'Regulator Reputation, Enforcement, and Environmental Compliance', Journal of Environmental Economics and Management, 50, pp. 519–40.
- Kazumasu Aoki and John W. Cioffi (1999), 'Poles Apart: Industrial Waste Management Regulation and Enforcement in the United States and Japan', Law and Policy, 21, pp. 213–45
- https://www.fca.org.uk/publication/occasional-papers/op16-24.pdf

Organisational measures to support everyday use of toolkits + Use of toolkits require endorsement by senior staff

- Errida A, Lotfi B. The determinants of organisational change management success: Literature review and case study. International Journal of Engineering Business Management. 2021;13. doi:10.1177/18479790211016273

"Move fast and break things" mentality

- Birkinshaw, Julian. (2022). Move fast and break things: Reassessing IB research in the light of the digital revolution. Global Strategy Journal. 12. 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. New York: Little, Brown, and Company, 2017. 321 pp. Figures, notes, index. Cloth, $19.72. ISBN: 978-0-316-27577-4. Business History Review. 2018;92(1):191-193. doi:10.1017/S000768051800020X

Certification and Liability

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

Ticking the box exercises

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

Greenwashing

- Mutua K, Powell-Turner J, Spiers M, Callaghan J. An In-Depth Analysis of Barriers to Corporate Sustainability. Administrative Sciences. 2025; 15(5):161. https://doi.org/10.3390/admsci15050161
- Free, C., Jones, S. and Tremblay, M.-S. (2024), "Greenwashing and sustainability assurance: a review and call for future research", Journal of Accounting Literature, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JAL-11-2023-0201

Role of procurement in toolkits adoption

- Howe, J. (2016). The regulatory impact of using public procurement to promote better labour standards in corporate supply chains. In Fair Trade, Corporate Accountability and Beyond (pp. 329-347). Routledge.
- https://www.oecd.org/en/topics/public-procurement.html
- Oishee Kundu, Elvira Uyarra, Raquel Ortega-Argiles, Mayra M Tirado, Tasos Kitsos, Pei-Yu Yuan, Impacts of policy-driven public procurement: a methodological review, Science and Public Policy, Volume 52, Issue 1, February 2025, Pages 50–64, https://doi.org/10.1093/scipol/scae058

Facebook's "like" feature and psychological effect on teens

- https://www.rochester.edu/newscenter/getting-fewer-likes-on-social-media-can-make-teens-anxious-and-depressed-453482/
- https://www.apa.org/monitor/2023/09/protecting-teens-on-social-media
- Translation of abstract ethical principles into actionable, practical frameworks
- Ibáñez, J.C., Olmeda, M.V. Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. AI & Soc 37, 1663–1687 (2022). https://doi.org/10.1007/s00146-021-01267-0
- Zhou, J., Chen, F. AI ethics: from principles to practice. AI & Soc 38, 2693–2703 (2023). https://doi.org/10.1007/s00146-022-01602-z
- Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, Francisco Herrera, Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation, Information Fusion, Volume 99, 2023, 101896, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2023.101896.
- Regulatory flexibility vs. regulatory uncertainty /Principles and uncertainty
- Black, Julia. 2008. Forms and paradoxes of principles-based regulation. Capital Markets Law Journal 3: 425–457
- Cunningham, Lawrence A. 2007. A prescription to retire the rhetoric of "Principles-based systems" in corporate law, securities regulation, and accounting. Vanderbilt Law Review 60: 1411–1493.
- Carter, R.B., Marchant, G.E. (2011). Principles-Based Regulation and Emerging Technology. In: Marchant, G., Allenby, B., Herkert, J. (eds) The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight. The International Library of Ethics, Law and Technology, vol 7. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1356-7_10

Regulatory Sandboxes

- https://www.fca.org.uk/publication/fca/fca-regulatory-sandbox-guide.pdf
- Ranchordas, Sofia and Vinci, Valeria, Regulatory Sandboxes and Innovation-friendly Regulation : Between Collaboration and Capture (January 16, 2024). Sofia Ranchordas & Valeria Vinci, Regulatory Sandboxes and Innovation-friendly Regulation: Between Collaboration and Capture, Italian Journal of Public Law, Vol. 1/2024, Forthcoming , Tilburg Law School Research Paper, Available at SSRN: https://ssrn.com/abstract=4696442 or http://dx.doi.org/10.2139/ssrn.4696442

- OECD (2023), "Regulatory sandboxes in artificial intelligence", OECD Digital Economy Papers, No. 356, OECD Publishing, Paris, https://doi.org/10.1787/8f80a0e6-en.
- Ranchordas, Sofia, Experimental Regulations for AI: Sandboxes for Morals and Mores (May 4, 2021). University of Groningen Faculty of Law Research Paper No. 7/2021, Available at SSRN: https://ssrn.com/abstract=3839744 or http://dx.doi.org/10.2139/ssrn.3839744
- Truby J, Brown RD, Ibrahim IA, Parellada OC. A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications. European Journal of Risk Regulation. 2022;13(2):270-294. doi:10.1017/err.2021.52

Co-Regulation

- Roger Clarke, Regulatory alternatives for AI, Computer Law & Security Review, Volume 35, Issue 4, 2019, Pages 398-409, ISSN 2212-473X, https://doi.org/10.1016/j.clsr.2019.04.008.
- Eijlander, Philip, Possibilities and Constraints in the Use of Self-Regulation and Co-Regulation in Legislative Policy: Experiences in the Netherlands - Lessons to Be Learned for the EU?. European Journal of Comparative Law, Vol. 9, No. 1, January 2005, Available at SSRN: https://ssrn.com/abstract=959148
- Hirsch, D. D. (2011). The law and policy of online privacy: regulation, self-regulation, or co-regulation. Seattle University Law Review, 34(2), 439-480
- Singapore's case
- Allen JG, Loo J, Campoverde JLL. Governing intelligence: Singapore's evolving AI governance framework. Cambridge Forum on AI: Law and Governance. 2025;1:e12. doi:10.1017/cfl.2024.12
- Lim, S. S., & Chng, G. (2024). Verifying AI: will Singapore's experiment with AI governance set the benchmark? Communication Research and Practice, 10(3), 297–306. https://doi.org/10.1080/22041451.2024.2346416
- Remolina, Nydia, AI Governance and Algorithmic Auditing in Financial Institutions: Lessons From Singapore (March 31, 2025). Singapore Management University School of Law Research Paper (forthcoming), SMU Centre for Digital Law Research Paper (forthcoming), Available at SSRN: https://ssrn.com/abstract=5199968 or http://dx.doi.org/10.2139/ssrn.5199968

Learning from failures

- Black, Julia, Learning from Regulatory Disasters (November 6, 2014). LSE Legal Studies Working Paper No. 24/2014, Available at SSRN: https://ssrn.com/abstract=2519934 or http://dx.doi.org/10.2139/ssrn.2519934

Intersectionality, Discrimination and Large Language Models/Need for AI literacy

- Bentley, C., Ramchurn, S. D., Parisio, I., McStay, A., & Oman, S. (2025, Apr 9). Responsible AI response to the DSIT's open call for evidence on the Digital Inclusion Action Plan. King's College London. https://doi.org/10.18742/PUB01-216

Need for interdisciplinary approaches

- Lu, C. Rethinking artificial intelligence from the perspective of interdisciplinary knowledge production. AI & Soc 39, 3059–3060 (2024). https://doi.org/10.1007/s00146-023-01839-2
- https://www.kcl.ac.uk/beyond-silos-why-ai-regulation-calls-for-an-interdisciplinary-approach
- Jonas Ammeling, Marc Aubreville, Alexis Fritz, Angelika Kießig, Sebastian Krügel, Matthias Uhl, An interdisciplinary perspective on AI-supported decision making in medicine, Technology in Society, Volume 81, 2025, 102791, ISSN 0160-791X, https://doi.org/10.1016/j.techsoc.2024.102791.
- Hine, C., & Barnaghi, P. (2024). Ethics and Artificial Intelligence in the Interdisciplinary Collaborations of Smart Care. Science, Technology, & Human Values, 0(0). https://doi.org/10.1177/01622439241302519
- Confiance.ai (2024) Towards the engineering of trustworthy AI applications for critical systems, Second Edition. Available: https://www.confiance.ai/the-confiance-ai-programme-publishes-its-second-white-paper/

Supply chain complexities and liability

- Beatriz Botero Arcila. AI Liability Along the Value Chain. MOZILLA. 2025, pp.68. ⟨hal-05025891⟩. https://sciencespo.hal.science/hal-05025891/
- Custers, B., Lahmann, H. & Scott, B.I. From liability gaps to liability overlaps: shared responsibilities and fiduciary duties in AI and other complex technologies. AI & Soc 40, 4035–4050 (2025). https://doi.org/10.1007/s00146-024-02137-1
- Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. Big Data & Society, 10(1). https://doi.org/10.1177/20539517231177620 (Original work published 2023)
- Clear rules can set expectations
- Winston J. Maxwell, Principles-based regulation of personal data: the case of 'fair processing', International Data Privacy Law, Volume 5, Issue 3, August 2015, Pages 205–216, https://doi.org/10.1093/idpl/ipv013

Asymmetry of power between BigTech companies and other market players, and regulators

- https://www.mediareform.org.uk/blog/is-big-tech-too-big-to-regulate-lse-event
- https://hai.stanford.edu/news/the-tech-coup-a-new-book-shows-how-the-unchecked-power-of-companies-is-destabilizing-governance
- Manganelli, A., Nicita, A. (2022). Regulating Big Techs and Their Economic Power. In: Regulating Digital Markets. Palgrave Studies in Institutions, Economics and Law. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-89388-0_6
- Lindman, J., Makinen, J., & Kasanen, E. (2023). Big Tech's power, political corporate social responsibility and regulation. Journal of Information Technology, 38(2), 144-159. https://doi.org/10.1177/02683962221113596 (Original work published 2023)

Authors and contributors

Writing team

RAi UK

Aled Lloyd Owen
CoS, RAi UK
University of Southampton

Sarvapali (Gopal) Ramchurn
CEO RAi UK
University of Southampton

Prokar Dasgupta
Chair, RAi UK Health & Social Care Working Group
KCL

Shuang Ao
University of Southampton

Pepita Barnard
University of Nottingham

Jayati Deshmukh
University of Southampton

Isabela Parisio
King's College London

Lokesh Singh
University of Southampton

Adarsh Valoor
University of Southampton

Maria Waheed
University of Nottingham

European Trustworthy AI Association

Bertrand Braunschweig

Karla Quintero

Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)

Maximilian Poretschkin

Ben Hawes
University of Southampton
(Freelance)

Foutse Khomh
Polytechnique Montréal

Workshop organisers

RAi UK

Sarvapali (Gopal) Ramchurn
CEO RAi UK
University of Southampton

Aled Lloyd Owen
CoS, RAi UK
University of Southampton

European Trustworthy AI Association

Nicolas Rebierre

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS

Maximilian Poretschkin

DOI : https://doi.org/10.5258/RAi/003

Participants

Shuang Au, University of Southampton
Vidhu Aul, KCL
Pepita Barnard, University of Nottingham
Amel Bennaceur, Open University
Sundeep Bhandari, NPL
Bertrand Braunschweig, IRT SystemX
Nick Bryan-Kinns, University of the Arts, London
Julien Chiaroni, SAFENAI
Cedric Couette, Safran
Prokar Dasgupta, KCL
Jayati Desmukh, University of Southampton
Kate Devlin, KCL
Farzana Dudhwala, Meta
Joel Fischer, University of Nottingham
Athina Georgara, University of Southampton
Ben Hawes, University of Southampton
Cari Hyde-Vaamonde, KCL
Matt Jones, Swansea University
Emma Kallina, University of Cambridge
Foutse Khomh, Polytechnique Montréal
Lars Kunze, UWE
Derek McAuley, University of Nottingham
John McDermid, University of York
Gina Neff, QMUL
Aled Lloyd Owen, University of Southampton
Isabela Parisio, KCL
Khalid Parkar, University of Southampton
Vikrant Patel, Sagacité
Colin Paterson, University of York
Angela Paul, Northumbria University
Sven Peets, Harper Adams University
Virginia Portillo, University of Nottingham
Maximilian Poretschkin, Fraunhofer IAID
Rob Procter, University of Warwick
Karla Quintero, IRT SystemX
Sarvapali (Gopal) Ramchurn, University of Southampton
Nicolas Rebierre, IRT SystemX
Lokesh Singh, University of Southampton
Alexandra Smyth, RAEng
Jack Stilgoe, UCL
Andrew Thompson, NPL
Adarsh Valoor, University of Southampton
Lucy Veale, University of Nottingham
Maria Waheed, University of Nottingham
Angela Westley, University of Southampton
Jennifer Williams, University of Southampton

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.

About the European Trustworthy AI Association

In a global context where artificial intelligence displays tremendous potential to transform industrial products, services and processes, the European Trustworthy AI Association is positioned as the driving force behind an ambitious European strategy for industrial and responsible AI. Its aim is to propel Europe to the forefront of innovation in trustworthy AI, by making our methodologies and tools an international benchmark.

The Association's mission is to transform these ambitions into concrete action, by creating an environment conducive to the emergence of innovative and reliable solutions, while ensuring regulatory and ethical compliance.

https://www.trustworthy-ai-association.eu/

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].