Assuring growth: Making the UK a global leader in AI assurance technology

BRIEFING PAPER

July 2024

SMF Social Market Foundation

By Jam Kraprayoon and Bill Anderson-Samways

Artificial intelligence could boost the economy and public sector productivity, but for these benefits to materialise concerns around the safety and reliability of AI tools need to be allayed. This report looks at the rising demand for AI assurance technologies and sets out how the UK can get a head start in this market.

KEY POINTS

- Artificial Intelligence (AI) could bring economic benefits worth trillions of pounds and transform UK public services.
- For those benefits to be realised, users need assurance that AI tools are safe, secure, and reliable. As a result, there is a burgeoning demand for new AI assurance technologies (AIATs) – solutions for AI alignment, security, auditing, authentication, and risk management.
- New modelling suggests the global AIAT market will reach $276 billion by 2030, but most AIATs currently need R&D investment to operate at scale.
- The UK has a head-start in the AIAT landscape, with a world-leading AI Safety Institute (AISI) and a fledgling AIAT start-ups scene. But uncertainties around regulation, technical priorities and demand mean that the global AIAT market is still very much up for grabs.
- Handled correctly, the UK could capture a massive share of the emerging AIAT market, complementing growth in the UK AI market more generally and supporting national security and regional industrial strategy.

RECOMMENDATIONS

- Announce a market shaping programme mobilising public and private sector investment to supercharge the UK's AIAT industry.
- Establish an inter-departmental group (IDG) to guide AIAT policy.
- Come up with a three-year roadmap for introducing mandatory AI assurance standards, including an MOU with the EU and US.
- Direct £10m of AISI's budget into seed funding for priority AIATs, and grant companies developing said AIATs free access to public compute.
- Commission a study analysing the technical and market barriers to developing and commercialising priority AIATs.
- Invest £50m in "pull mechanisms" (pay-outs contingent on achieving specific technological goals, such as prizes and milestone payments).

UKDayOne

IAPS Institute for AI Policy and Strategy

ABOUT THIS BRIEFING

This briefing was has been published simultaneously by UK Day One Project and the Social Market Foundation. The authors are researchers at the Institute for AI Policy and Strategy (IAPS).

The UK Day One Project is helping the new government with implementation-ready policies for growth and progress in the UK.

The Institute for AI Policy and Strategy (IAPS) is a think tank working to understand and navigate the transformative potential of advanced AI. Its mission is to identify and promote strategies that maximize the benefits of AI for society and develop thoughtful solutions to minimize its risks.

ABOUT THE AUTHORS

Jam Kraprayoon is a Researcher on the Policy & Standards team at the Institute for AI Policy and Strategy (IAPS). Prior to IAPS, he was a manager at the Effective Institutions Project and a program officer at the Asian Productivity Organization, where he focused on public sector reform, regulation, and strategic foresight. He holds an MPhil in Politics from the University of Oxford and a BSc in Government from the London School of Economics.

Bill Anderson-Samways is a Research Analyst on the Policy & Standards team at the Institute for AI Policy and Strategy (IAPS). He was previously a researcher at the Social Market Foundation where he focused on climate policy. He holds an MPhil in Anthropocene Studies and a BA in Politics and International Relations, both from the University of Cambridge.

CHALLENGE AND OPPORTUNITY

AI could generate an extra £15 trillion for the global economy by 2030.[1] AI also stands to transform public services - the UK public sector could make savings of up to £40 billion a year just by embracing existing AI tools.[2]

However, the benefits from AI will not be realised unless the technology is safe, secure, and reliable. In healthcare, for example, failures in AI-assisted diagnosis could increase existing public mistrust, inhibiting NHS AI adoption.[3] Similarly, businesses will be reluctant to adopt AI systems if they cannot be trusted to perform tasks as intended. The autonomous vehicle (AV) industry is instructive: in the United States, 93% of the public have concerns about AVs, with safety coming in at number one, according to Forbes research.[4] Forbes notes that these safety concerns are a major reason for slow growth projections in the AV market.

Given the risks tied to the opacity and unreliability of advanced AI systems, both AI developers and companies planning to deploy AI-based services stand to gain from effective risk management. Such risk management will minimise losses from damaged assets, service disruptions, insurance costs, mistrust among end-users, and litigations from actual harms.

The technologies needed to ensure AI systems are robust are **AI assurance technologies (AIATs)**.[5] AIATs are the software, hardware, and services that enable organisations to better manage risks from AI (Table 1 lists five major AIAT areas, along with specific examples). Recent estimates suggest that a global AIAT market could reach $276 billion by 2030.[6] However, AIATs do not yet exist at scale.

Table 1: Major AI assurance technology areas

AI Assurance Technology Areas	Description
AI alignment technologies	Techniques to align AI systems with the goals of their developers. Example: Automated alignment research, i.e. using existing advanced AI systems to aid in the development of alignment techniques.[7]
Security solutions for AI systems and infrastructure	Technologies to protect AI systems from unauthorised access or disruption, including both hardware security and cybersecurity. Example: Hardware-integrated monitoring mechanisms.[8]
Tools for auditing AI systems	Tools for evaluating AI risk, as well as platforms for ensuring compliance with regulations. Example: Mechanistic interpretability tools.[9]
Tools for managing AI risk	Tools to formalise government, third-party or corporate oversight of the training and deployment of AI systems. Example: Tools or APIs to provide structured access to (components of) models (e.g. model weights).[10]

Digital authentication tools for AI systems	Tools for identifying and marking AI-generated content and AI-enabled agents. Example: Agent IDs.[11]

The UK has an opportunity to become a global leader in AI assurance technology, and to capture a disproportionate share of the coming AIAT market. Indeed, the potential to develop *safe* AI systems is perhaps the UK's major comparative advantage in the global AI market. The UK possesses the world's first and largest AI Safety Institute (AISI), as well as pools of world-leading AI safety talent in organisations such as the Alan Turing Institute and Google DeepMind.

Already, large UK-based private sector investors like Entrepreneur First have announced programmes to promote AIATs.[12] The UK also possesses numerous start-ups which specialise in the early-stage AIATs, such as Aligned AI,[13] Mindgard,[14] Advai[15] and Holistic AI.[16] In addition, Deloitte recently acquired Gryphon Scientific, a company working with leading AI companies (OpenAI and Anthropic) on chemical, biological, radiological and nuclear red-teaming, evaluations, and other assessments.[17] All of this proves that UK companies can establish a firm niche in this space.

These developments give the UK a head start. However, potential AIAT companies face three major challenges in finding success in this market:

1. Uncertainties about the future regulatory environment around AI, both domestically and internationally, making it harder for companies to set R&D priorities for AIAT.
2. Uncertainties about key opportunities in the AIAT space, e.g. what assurance solutions will be most useful for AI developers, operators of AI infrastructure, and businesses looking to capitalise on AI-led growth.
3. Risks associated with investing in R&D for specific technologies where demand is presently weak.

With the right policies in place, the UK government can address these challenges, resulting in far more private investment flowing into this sector and more start-ups springing up in the UK. The exact return is, of course, difficult to predict. If the UK were to capture a similar share of the above-mentioned $276 billion figure as its 2022 share of the global cybersecurity market (around 8.5%),[18] **UK companies could be earning almost $24 billion (£18 billion) in annual revenue by 2030. However, as suggested, the AIAT market is the UK's for the taking. It's possible that the country could capture a significantly higher fraction of that $276 billion market.**

While it is likely that many of the gains would benefit London and the South East, there could also be opportunities for other parts of the UK. The Bristol and Bath region, for example, contains a large number of cybersecurity firms.[19] Investment in cybersecurity tools for AI systems in that region could lead to significant growth, bolstering the West of England Combined Authority's plan to create a "Western powerhouse" in the region.

Developing safe and reliable AI systems will also be vital to ensure UK national security. AI systems with security vulnerabilities could be stolen by criminals, terrorist actors, or rogue states, who could then use such systems to directly attack the UK and allies (for example with AI-powered cyber attacks) or illicitly acquire resources (for example via mass spear phishing campaigns).[20] Meanwhile, the UK Ministry of Defence (MOD) aims to become the world's most trusted defence organisation where it comes to AI.[21] This goal will not be realised, however, if MOD AI systems experience unexpected robustness failures which lead to casualties among defence personnel or civilians.

PLAN OF ACTION

To kickstart the development of an AIAT industry in the UK, the new Secretary of State for Science, Innovation and Technology should announce a market-shaping programme focused on mobilising public and private-sector investment into key AIATs. The programme will be designed by a specialised team formed within the Department for Science, Innovation and Technology's (DSIT) AI Policy Directorate (see recommendation 5, below). A well-publicised announcement will signal to businesses, investors, and researchers that the government regards AIAT as a priority, thereby boosting investment confidence.

The following recommendations outline the major elements of that market-shaping programme.

Recommendation 1: The Secretary of State for Science, Innovation and Technology should establish an inter-departmental group (IDG) to develop high-level strategy around AIAT innovation. The IDG should consist of DSIT's AI Safety Institute (AISI) and AI Policy Directorate, alongside the Department for Business and Trade (DBT). The IDG should outline specific AIAT development priorities, informed by the technical expertise of AISI. While this group should ultimately be responsible for deciding what specific AIAT is prioritised, in the FAQ we outline specific technologies that should be strongly considered.

Recommendation 2: The government should establish a three-year roadmap for introducing mandatory standards around AI risk assessment, safety and security. (Our FAQ outlines what mandatory standards could look like). In the meantime, DSIT should enhance the effectiveness of existing standards by introducing new policies on international harmonisation and public procurement.

Mandatory standards are, at base, a necessity for ensuring safety and security. However, that very fact will benefit AI developers in the UK:

- As outlined above, current AI systems have significant reliability and security issues. *The Economist* notes a major discrepancy between investor confidence in AI systems and actual revenues estimated for this year.[22] The core reason here is slow adoption, in large part due to "concerns about data security, biased algorithms and hallucinations".
- Standards addressing such concerns would give the private and public sectors more confidence in integrating AI into their business, increasing the rate of adoption.

- Standards need only be mandatory in the case of significant risks - for example, risks arising from the use of AI in critical national infrastructure.

Standards will also drive growth in the AIAT market itself:

- Compliance requirements will drive demand for AIAT among firms, much as national emissions standards have driven large-scale investment in UK climate tech. In the AI sphere, the EU AI Act is an already-existing source of demand for AIAT.[23] Article 55, for example, obliges providers of risky advanced AI systems to perform evaluations, mitigate risks, document safety incidents, and ensure adequate cybersecurity protections, which all demand specific AIAT solutions.[24] As outlined in our FAQ, existing UK voluntary standards are significantly more detailed and technically adept than EU (and US) law.[25] Making those standards mandatory (once they are well-developed enough) will drive demand for many more kinds of AIAT.
- Companies offering AI-based services will have an easier time planning around a publicly announced roadmap, with clear communication channels set up between government and the private sector.
- Companies' R&D priorities in the AIAT space will benefit from a clear understanding of the government's own AIAT priorities.

As part of this roadmap, the Secretary of State for Science, Innovation and Technology should establish a Memorandum of Understanding (MOU) between DSIT, the US Department of Commerce, and the EU AI Office, to harmonise on standards around AI evaluations, safety and security. Harmonisation is critical because EU and US standards will strongly influence the demand for the kinds of AIAT needed by industry. For example, UK recognition under Article 39 of the EU AI Act[26] will enable UK organisations to conduct assessments of 'high-risk AI systems' in the EU, provided they meet specified criteria.

There is currently an MOU between the US and UK AI Safety Institutes, but this does not include the EU AI Office.[27] In addition, the US and UK AISIs are not regulatory bodies. Given that some standards will become regulations, DSIT's AI Policy Directorate as well as the broader US Department of Commerce (in which the US AISI is located) should be included as well.

AIAT standards should also be integrated into public sector procurement policies due to the critical nature of government services. Higher safety and security standards for AI in public sector procurement are essential to prevent disruptions that could have severe consequences for national infrastructure and security.

As AI is a fast-moving field, it is important to note that mandatory standards should be reviewed frequently, to ensure that they keep up with the latest technical developments. The IDG could perform this review function.

Recommendation 3: Following the IDG's high-level strategy, AISI should direct £10 million of its already allocated budget towards seed grants to fund R&D into priority AIAT solutions. It should also grant free public compute to AIAT researchers and companies. These grants should focus on promising, earlier stage AIATs that need additional validation before being ready to prototype or become a product.

Recommendation 4: DSIT's AI Policy Directorate should commission a study analysing the technical and market barriers to developing and commercialising priority AIATs, to establish which pull mechanisms are best suited to different technologies.

Pull mechanisms (such as prizes, advance market commitments, and milestone payments) incentivise innovation by only rewarding entities *after* they meet specific goals, i.e. after projects solve the specific AI assurance challenges set out by DSIT. Pull mechanisms will create early momentum for private investment while broader market demand is still solidifying.

This study should include an assessment of the comparative advantages of the UK in the AIAT market relative to major international competitors (e.g. the EU, the US, China, Japan, South Korea).

Recommendation 5: DSIT should invest an initial £50 million in developing market-shaping pull mechanisms for researchers and companies that develop AIATs or solve key AIAT innovation challenges. DSIT should seek support from private donors to match public investment for these pull mechanisms. Based on outcomes of the initial round, DSIT should consider additional investments based on technological maturity and the commercial prospects of related AIAT firms.

To implement the pull mechanisms programme, DSIT's AI Policy Directorate should:

1. Establish a team of AI assurance and safety specialists, economists, and contracting specialists to develop and administer AIAT-specific pull mechanisms. This team should draw expertise from other government organisations such as AISI and DBT, as well as from civil society.
    a. The team should also act as a touchpoint for the UK AIAT ecosystem, connecting firms to government in order to communicate details around AIAT priorities, the standard-setting roadmap, and the market-shaping programme. The team could meet directly with interested companies and organise roundtable sessions that bring together AI companies, potential AIAT firms, and key policymakers.
2. Partner with private donors, including philanthropists and investors, to match public funding for these mechanisms. For example, for an incentive prize of £1 million, a private donor could provide a minimum of £500,000 to match this investment.

One example of a potentially promising pull mechanism would be a £10 million prize to develop privacy-preserving machine learning techniques for model assurance and forensics (under the AIAT category "tools for auditing AI systems"). These techniques would enable model developers or outside auditors to prove that certain characteristics of a model are true without violating privacy, for example that it was developed with certain safeguards in place or that it does not contain sensitive information in its training dataset.

FAQs

Why is government involvement needed to support/develop the AIAT industry? Why can't this be done by the private sector alone?

As mentioned in the main memo text, firms face three main challenges when thinking about whether to invest in the AIAT space: regulatory uncertainty, uncertainties in AIAT opportunity identification, and risks from investing in R&D before demand is established. The AIAT sector has significant long-term growth potential and is likely to bring social and national security benefits, but targeted market stimulation is needed to ensure the UK takes an early lead. Without targeted industrial strategy, companies will find it difficult to keep up in a fast-changing industry.

Nonetheless, the private sector has a large role to play to make the AIAT sector successful and the role of the government is to enable this. This plan is designed to keep industry and government in close coordination. First, DSIT and AISI should be in communication with leading AI developers and infrastructure providers to understand their AIAT needs. Second, pull mechanisms developed by the government should be priced correctly to incentivize firms to enter the AIAT space. Third, the government can leverage the expertise of civil society and private investors by identifying which demand-pull mechanisms they are willing to co-fund.

Will mandatory standards for AI safety and security slow down innovation?

Overly restrictive regulations could harm AI's potential for developing economic growth and benefits to research and innovation. However, the need for standards around safety and security also reflect fundamental features of current frontier AI systems. AI systems can be misaligned and unreliable, and they can be misused to cause serious political, economic and social harms. AI systems are also vulnerable to novel forms of attacks, e.g. data poisoning and adversarial examples.

Standards, alongside supporting technologies, are needed to manage these risks and allow AI to be deployed in a way that leads to secure and stable growth. Standards need only be mandatory when risks are significant, and when risks are significant, innovation can only happen when there is assurance of safety and security. Without these standards, both the private and public sector are likely to be more hesitant to leverage AI, which will slow down innovation.

What should the mandatory standards on AI safety and security mentioned in the three-year roadmap cover?

Mandatory standards could cover the key elements of DSIT's emerging processes for frontier AI safety agenda, including responsible capability scaling, model evaluations and red teaming, model reporting and information sharing, and security controls including securing model weights (among other areas).[28]

DSIT's emerging processes agenda is far more detailed and comprehensive than AI regulations in the US and EU. In the US, Executive Order 14110 focuses mainly on reporting requirements.[29] That is a good start, but it is not sufficient - as an analogy,

if climate change regulations had been restricted to emissions reporting requirements, they would not have driven anything like the necessary innovation in clean technology. The EU AI Act (Article 55) is broader, touching on model evaluations, risk-mitigation, and security controls.[30] However, it is much less detailed than DSIT's agenda, perhaps because DSIT possesses significantly more technical expertise at present. DSIT's agenda is moreover the only document to outline processes for responsible capability scaling (which IAPS research has addressed elsewhere).[31]

However, none of the standards in DSIT's agenda are yet mandatory. That is because they are still a work in progress, and are being continually updated until they are regulation-ready. Thus, the exact content of binding AI standards will depend on which of DSIT's standards are well-developed enough to be translated into legislation within the next three years. The next government should thus continue to prioritise the development of DSIT's standards, while committing to putting those standards on a statutory footing as soon as possible.

How does AI Assurance Technology differ from DSIT's existing "Portfolio of AI assurance techniques"?

We drew our definition of AIAT from a recently-published report by Juniper Ventures.[32] That definition is quite distinct from DSIT's existing portfolio of AI assurance techniques.[33]

DSIT's portfolio is essentially an options list of techniques that *already* exist, based on past case-studies, whereas AIAT covers technologies that do not yet exist or are still in their infancy. The latter is a far larger bucket, given that AI assurance is still a nascent field. AIAT therefore has massive growth potential, with Juniper Venture's modelling suggesting that the AIAT industry could grow from $1.6 billion in 2023 to $276 billion in 2030.[34] DSIT's portfolio, on the other hand, was not designed to be a list of growth areas.

A non-exhaustive list of AIATs, and our categorisations of them, can be seen under the question "What specific technologies should the government prioritise in the AI assurance technology space?", below. As you can see, it differs substantially from those employed in DSIT's portfolio.

Why is this policy package a good use of government money?

Outside the substantive reasons laid out in the main briefing text about why the UK government should make early investments in the AIAT sector, this policy package has been designed to be lean and efficient in terms of spending.

Focusing on demand-pull mechanisms, as we do in recommendations 4 & 5, means that while funding is committed earlier, payouts are outcomes-based, i.e. payment is made only when a core AIAT challenge has been met by a firm or research group. Pull mechanisms are meant to stimulate the market and draw in private sector investment by creating a clear initial demand. For example, NASA's Lunar Lander Challenge in the US awarded $2 million in prize funding but spurred a subsequent $20 million total investment.

In addition, we recommend that the government actively seek matching funds from private philanthropists and investors. Given AIAT's potentially significant wider social impact combined with their market potential, it seems feasible to create pull mechanisms that are funded jointly by the public and private sector. Public-private funding for pull mechanisms has been used very successfully in the vaccine space, for example with GAVI Alliance's Pnuemococcal Vacine AMC or the Medicines for Malaria Venture.

It should be noted that this set of policy is focused on R&D, which counts as investment under most political parties' treasury spending rules and can therefore be financed by borrowing, unlike spending more generally.

How can the UK government ensure that DSIT spending on AIAT solutions doesn't end up going to large industry players, rather than scaleups/startups?

In terms of our recommendation regarding AISI allocating £10 million in seed funding to companies working on AIAT solutions, it is in principle possible to make grants easier to access for small and medium enterprises (SMEs). As a precedent, Innovate UK's Smart Grants programme contains favourable conditions for SMEs - for example, each project which applies for a grant must contain at least one SME.[35] Whether or not that is the correct approach in this particular case, however, is up to policymakers. Ultimately the priority must be to deliver innovative AIAT solutions.

In terms of our recommendation regarding the £50 million DSIT pull mechanism program, it is not possible to guarantee that the spending ends up going to SMEs. Pull mechanisms, by nature, cannot determine in advance which kinds of businesses receive funding. Instead, funding goes to whichever company first develops the technological solution in question.

However, regardless of the exact shape of different policy mechanisms, we do expect that SMEs will be more likely than large companies to apply and compete for AIAT funding, because:

- The fledgling AIAT space at present mostly consists of startups, not large companies (such as Aligned AI,[36] Mindgard,[37] Advai[38] and Holistic AI[39]).
- AIAT solutions often require a strong degree of technical expertise, which can often only come from start-ups which have spun out of university departments. For example, Aligned AI was founded by an Oxford Professor, Stuart Armstrong.[40]
- Most AIAT research is basic research involving uncertain pay-offs. We therefore expect that large companies like Deloitte will often be unwilling to make bets on AIAT, as it represents a significant opportunity cost against their regular, more reliable business streams. Start-ups, on the other hand, can only scale by making risky bets, so they don't face a comparable opportunity cost.

Given that AI companies are globally mobile, could there be an unintended deterrent for AI companies created by mandating standards around AI risk assessment?

As stated in the report, we expect that mandatory AI standards will create business for AI companies, rather than hurting it. Reliable, safe AI tools are more likely to be adopted by businesses and the public sector.

In addition, we advocate the development of a "roadmap" towards mandatory standards, so that firms have ample time to adjust to future regulatory requirements. We expect that this will introduce greater regulatory certainty than in other jurisdictions, which will attract AI companies to the UK.

Finally, it is worth noting that the most important overseas jurisdictions *already* have mandatory standards, for example the European Union's AI Act (see in particular Article 55)[41] and the Biden administration's Executive Order 14110 (see in particular the reporting requirements in section 4.2(a)).[42] UK standards will by no means necessarily be more *intense* than those in other jurisdictions. DSIT possesses much stronger technical expertise than either the EU or the US governments at present, so its regulations are, if anything, likely to be better targeted. For example, while Article 55 of the EU AI Act simply mandates that AI models possess "an adequate level of cybersecurity" - quite a broad-brush requirement - the UK's standards could be more specific, for example detailing a list of security measures that need to be in place.[43]

What specific technologies should the government prioritise in the AI assurance technology space?

As mentioned, AIAT is an emerging field, so the overleaf list is still largely illustrative. The ultimate decision regarding which AIATs should be prioritised should be made by the IDG, and may include technologies outside of this list.

Table 2: Examples of major AI assurance technology areas

AI Assurance Technology Areas	Description	Examples
AI alignment technologies	Techniques to align AI systems with the goals of their developers that have historically been neglected by industry, according to an upcoming paper from IAPS.[44]	Multi-agent safety, i.e. techniques to prevent failures arising from interactions between multiple AI systems (such as automated trading algorithms). Power-aversion (as defined in frameworks such as the MACHIAVELLI benchmark),[45] to mitigate risks from autonomous systems such as unintended resource-acquisition - e.g. a financial trading bot causing large-scale economic losses by conducting spear phishing campaigns without its manager's knowledge. Toy models of misalignment, i.e. creating simple systems with concerning behaviours to test proposed alignment techniques. Formal safety guarantees, i.e. the use of mathematical and philosophical formalisms to understand and mitigate risky properties of systems. UK ARIA's Safeguarded AI programme, currently funded to the tune of £59m, is an example of existing formal safety guarantees research; this programme could be expanded, or complementary smaller programmes could be established at ARIA.[46] Honesty and transparency, i.e. ensuring that systems accurately communicate the rationales behind their decisions. Automated alignment research, i.e. using existing advanced AI systems to aid in the development of alignment techniques.
Security solutions for AI systems and infrastructure	Security solutions to protect AI systems and infrastructure against unauthorised access or disruption. That includes hardware security, to prevent attacks on compute chipsets during training	Hardware-integrated logging and monitoring devices which can detect network activity for patterns indicating AI training runs, to guard against e.g. the theft of chips for the training of malicious models.[47] Tamper-proof device enclosures, i.e. physical enclosures to prevent chips from being stolen without compromising chip performance.[48] Techniques for confidential computing and homomorphic encryption, both of which enable models to be trained or fine-tuned without ever decrypting model weights.[49]

	runs. It also covers cybersecurity, both to prevent conventional cyberattacks that aim to e.g. steal model weights and to prevent novel forms of attacks such as data poisoning or adversarial examples.	Privacy-enhancing technologies (PETs) with Access Control Mechanisms. These enable cloud computing organisations to issue access permissions to specific companies or individuals, again reducing the risk that model weights will be stolen during training or fine-tuning. Improved data encryption tools for AI systems, which can scramble sensitive data so that it cannot be read during transmission.[50] AI Firewalls,[51] i.e. security devices which can validate model inputs and outputs transmitted via API. These can protect deployed AI models from prompt injection attacks.[52] Defensive AI, i.e. AI systems developed and deployed to protect businesses and defend against cybersecurity threats to critical infrastructure.[53]
Tools for auditing AI systems	Tools for evaluating AI risk, as well as platforms for ensuring compliance with regulations.	Improved techniques for identifying and mitigating algorithmic bias, for example fairness indicators,[54] what-if tools,[55] and data-labelling.[56] Adversarial robustness testing,[57] namely attempts to identify harmful behaviours in models using automated software tools[58] or human red-team assessments.[59] Tools and platforms for improved AI interpretability:[60] • Some interpretability tools do not look “inside” models, e.g. software which can automatically examine how changes in AI inputs affect outputs. • Meanwhile, mechanistic interpretability techniques such as activation patching attempt to open the “black box” of AI systems by understanding the concepts represented by internal model weights.[61] Risk evaluation tools are more specific to particular industries, e.g. healthcare, energy, or finance. Services for pre-deployment compliance auditing, aiming to identify vulnerabilities in AI systems to ensure that companies are not in violation of regulations before they bring their products to market. Some platforms use automated tools here.[62] Services for auditing conformity with data regulations. Data used during training, inference, and fine-tuning may be subject to regulatory requirements regarding security,

		quality, or anonymisation. Data Protection Impact Assessments are one useful tool to test compliance here.[63] Services for auditing hardware conformity, for example onsite inspections to test the security of AI chipsets or data centers, or procedures to validate whether protocols are in place to limit access to model weights. Services for auditing companies' governance procedures, for example regarding whistleblowing. Privacy preserving machine learning tools for model assurance and forensics.[64] These techniques would enable model developers or outside auditors to prove that certain characteristics of a model are true, for example that it was developed with certain safeguards in place or that it does not contain sensitive information in its training dataset.
Tools for managing AI risk	Tools to formalise government, third-party or corporate oversight of the training and deployment of AI systems.	Policy libraries and testing tools, which attempt to simulate extreme scenarios in order to test the robustness of company policies.[65] Software for managing reporting and regulatory conformity, using pre-created or tailored templates. Tools or APIs to provide structured access to (components of) models (e.g. model weights).[66] Software for AI observability, namely platforms which continuously collect and present data on model inputs, internal states, and decisions, facilitating the detection of (for example) unexpected behaviours by MLOps practitioners.[67] Software for monitoring AI infrastructure, which covers both infrastructure systems that use AI, and infrastructure systems necessary to support the use of AI (such as cloud computing). Such software monitors metrics such as usage patterns, hardware temperature, and resource consumption, to ensure that users are forewarned in the case of system derailment. Incident response software, namely platforms to help organisations identify and mitigate AI incidents.

Digital authentication tools for AI systems	Tools for identifying and marking AI-generated content and AI-enabled agents.	Digital signatures which insert identifying digital information into data files, enabling users to trace their origin.[68] Similarly, tools for forensic watermarking embed identifying information into media files.[69] Other software for the management and provenance tracking of digital assets, for example blockchain-based data provenance.[70] Just as there can be identification tools for AI-generated content, there could also be watermarks or IDs that identify AI agents to service providers or the general public.[71] Visual search software, which uses computer vision to identify similarities between media, to aid with (for example) detecting IP violations. Moderation software and services for AI-generated content.

ENDNOTES

[1] https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

[2] https://www.institute.global/insights/politics-and-governance/new-national-purpose-ai-promises-world-leading-future-of-britain

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7133482/#sec-a.d.gtitle

[4] https://www.forbes.com/advisor/legal/auto-accident/perception-of-self-driving-cars/

[5] https://www.aiat.report/report/AIAssureranceMarket/AI-Assurance-Tech-Market-Forecasts

[6] https://www.aiat.report/report/AIAssureranceMarket/AI-Assurance-Tech-Market-Forecasts

[7] https://arxiv.org/html/2406.01252v1

[8] https://www.cnas.org/publications/reports/secure-governable-chips

[9] https://www.anthropic.com/news/mapping-mind-language-model

[10] https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements

[11] https://arxiv.org/abs/2401.13138

[12] https://www.joinef.com/posts/introducing-def-acc-at-ef/

[13] https://buildaligned.ai/company

[14] https://mindgard.ai/

[15] https://www.advai.co.uk/

[16] https://www.holisticai.com/

[17] https://www2.deloitte.com/us/en/pages/consulting/solutions/biotech-consulting-and-data-analytics-solutions.html

[18] https://www.statista.com/statistics/1227894/revenue-of-cyber-security-sector-uk/; https://www.statista.com/outlook/tmo/cybersecurity/worldwide?currency=GBP

[19] https://techspark.co/cyber/

[20] https://www.cam.ac.uk/stories/malicious-ai-report

[21] https://assets.publishing.service.gov.uk/media/65bb75fa21f73f0014e0ba51/Defence_AI_Playbook.pdf

[22] https://www.economist.com/finance-and-economics/2024/07/02/what-happened-to-the-artificial-intelligence-revolution

[23] https://www.aiat.report/report/RiskManagement/Risk-forthcoming#unreliability-of-ai

[24] https://artificialintelligenceact.eu/article/55/

[25] https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety#responsible-capability-scaling

[26] https://artificialintelligenceact.eu/article/39

[27] https://www.gov.uk/government/news/uk-united-states-announce-partnership-on-science-of-ai-safety

[28] https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety#executive-summary

[29] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[30] https://artificialintelligenceact.eu/article/55/

[31] https://www.iaps.ai/research/responsible-scaling

[32] https://www.aiat.report/

[33] https://www.gov.uk/guidance/portfolio-of-ai-assurance-techniques

[34] https://www.aiat.report/

[35] https://www.ukri.org/councils/innovate-uk/guidance-for-applicants/guidance-for-specific-funds/smart-innovation-funding-guidance/

[36] https://buildaligned.ai/company

[37] https://mindgard.ai/

[38] https://www.advai.co.uk/

[39] https://www.holisticai.com/

[40] https://buildaligned.ai/company

[41] https://artificialintelligenceact.eu/article/55/

[42] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[43] https://artificialintelligenceact.eu/article/55/

[44] https://docs.google.com/document/d/1d676wl4-J71gjb1iK8lFkArRNMF_WaslU7jyt6eLEto/edit#heading=h.vt86d7f9uye6

[45] https://arxiv.org/abs/2404.09932

[46] https://www.aria.org.uk/programme-safeguarded-ai/

[47] https://www.iaps.ai/research/secure-governable-chips

[48] https://www.circuitinsight.com/programs/53605.html

[49] https://confidentialcomputing.io/2023/03/29/confidential-computing-and-homomorphic-encryption/

[50] https://link.springer.com/chapter/10.1007/978-3-030-68176-0_7

[51] https://www.robustintelligence.com/platform/ai-firewall-guardrails

[52] https://www.ibm.com/topics/prompt-injection

[53] https://blog.google/technology/safety-security/google-ai-cyber-defense-initiative/

[54] https://www.tensorflow.org/tfx/guide/fairness_indicators

[55] https://www.tensorflow.org/tensorboard/what_if_tool

[56] https://aws.amazon.com/what-is/data-labeling/

[57] https://developers.google.com/machine-learning/resources/adv-testing

[58] https://www.robustintelligence.com/what-is-algorithmic-ai-red-teaming

[59] https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/

[60] https://www.larksuite.com/en_us/topics/ai-glossary/interpretability-in-ai-and-why-does-it-matter

[61] https://rome.baulab.info/

[62] https://www.dekra.com/en/dekra-and-latticeflow-launch-ai-safety-assessment-services/

[63] https://www.mdpi.com/1999-5903/12/5/93

[64] https://ifp.org/where-can-federal-ai-rd-funding-go-the-furthest/

[65] https://www.preamble.com/solution

[66] https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements

[67] https://docs.dynatrace.com/docs/observe-and-explore/dynatrace-for-ai-observability

[68] https://www.researchgate.net/publication/233391380_A_survey_on_digital_signatures_and_its_applications

[69] https://support.cimediacloud.com/hc/en-us/articles/26068705038355-Forensic-Watermarking

[70] https://sol.sbc.org.br/index.php/wblockchain/article/view/12975

[71] https://arxiv.org/abs/2401.13138