The Alan Turing Institute

Centre for Emerging Technology and Security

RESEARCH REPORT

Securing the UK's AI Research Ecosystem

Megan Hughes, Sarah Mercer, Alexander Harris, Annie Benzie, Sam Williams and Elfreda Kenneison

March 2025

About CETaS

The Centre for Emerging Technology and Security (CETaS) is a research centre based at The Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to inform UK security policy through evidence-based, interdisciplinary research on emerging technology issues. Connect with CETaS at cetas.turing.ac.uk.

This research was supported by The Alan Turing Institute's Defence and Security Grand Challenge. All views expressed in this report are those of the authors, and do not necessarily represent the views of The Alan Turing Institute or any other organisation.

Acknowledgements

The authors are grateful to all those who took part in a research interview or workshop for this project, without whom the research would not have been possible. Thank you to Dr Dan Martin and Josh Collyer for their technical expertise and contributions to this report. The authors would also like to thank Dr Andrew Rogoyski and Joe C for their valuable feedback on an earlier version of this report.

Cite this work as: Megan Hughes, Sarah Mercer, Alexander Harris, Annie Benzie, Sam Williams and Elfreda Kenneison, "Securing the UK's AI Research Ecosystem," *CETaS Research Reports* (March 2025).

Executive Summary

AI is a valuable dual-use technology. The UK's world-leading AI research ecosystem is, therefore, a high-priority target for state threat actors seeking technological advantage. AI research develops rapidly, and early-stage academic research is often informal and collaborative. This creates opportunities for state threat actors to acquire knowledge or steal intellectual property (IP) with the intention of using UK-developed research and capabilities for malicious purposes. This threat requires an urgent, coordinated response from the UK Government and the higher education sector. To develop mitigations, they will need to strike a balance between the open nature of academic AI research and effective research security practices. This report explores the current constraints on AI research security and makes 13 recommendations for building the resilience of the UK's academic AI sector.

Key research findings

1. There is a **fundamental tension between academic freedom and research security**. This is compounded by a **lack of incentives for researchers to follow existing government guidance**.
2. Individual academics must often make personal judgements on the risks of their research. But **predicting the dual-use risks of early-stage research is very hard** and **awareness of the threat is not consistent** across the academic community.
3. **Culture change is needed within academia** to ensure research security is perceived as essential to high-quality research.
4. **The research security landscape is complex and diffuse,** and academics find existing procedures burdensome and constraining. **Clear information on the threat landscape and continued support in raising awareness** must be provided by the Government.
5. Short-term mitigations can create temporary friction for state threat actors, but a **long-term strategy to address structural problems such as funding gaps and talent retention** is needed to build a resilient AI research ecosystem.

Recommendations for the UK Government

1. The UK Government should commission a classified mapping of the AI higher education research ecosystem. This will provide the Government with a clearer overview of existing vulnerabilities and enable the provision of targeted support to institutions.

2. The Department for Science, Innovation and Technology (DSIT), with support from the National Protective Security Authority (NPSA), should provide regularly updated, direct guidance to research-intensive universities on international institutions deemed high-risk for funding agreements and collaborations. This will provide additional clarity to researchers.

3. The UK Government should provide DSIT with dedicated funding to grow the Research Collaboration Advice Team (RCAT). The RCAT is a key conduit of information between the Government and academia, and should be empowered to further invest in specialist technical staff and research capabilities to support academic due diligence.

4. DSIT should produce a white paper on the drivers of AI talent retention in academia. The AI Opportunities Unit should prioritise plugging the AI skills gap and encouraging young people into academic research roles to ensure domestic talent is retained and the UK remains at the forefront of high-quality technological research and development (R&D).

5. The NPSA and the National Cyber Security Centre (NCSC) should engage more widely with UK-based publishing houses, academic journals, and other research bodies to brief senior decision-makers on the threat landscape and offer tailored support to develop research security-minded policies.

6. The NPSA should declassify and publish case studies of relevant threats that have been intercepted or disrupted.

7. UK Research and Innovation (UKRI) should provide specific grant funding opportunities for research security activities to encourage institutions to invest in research security training and supporting infrastructure.

8. UKRI Grants Standard Terms and Conditions (T&Cs) should provide clarity to researchers on the guidance and legal provisions to which they must adhere. The T&Cs should explicitly outline unacceptable behaviours and require mandatory auditing of security practices to incentivise best practice.

9. UKRI should set up a Research Security Committee to report directly to its board on state threats and risk mitigations and to serve as an audit function for research security-related T&Cs.

Recommendations for academia

10. All academic institutions should be required to deliver research security training (based on Trusted Research guidance) to new staff and postgraduate research students as a prerequisite for grant funding. Training should be accredited by the NPSA.

11. The academic sector should develop a centralised due diligence repository to document risks and inform decision-making on AI research partnerships and collaboration. This repository should be hosted by a trusted partner, such as Universities UK (UUK) or UKRI.

12. Research-intensive universities should set up research security scrutiny committees to support AI (and other critical technology) researchers in their risk assessments.

13. Pre-publication risk assessment for AI research should be standardised across major AI journals and academic publishing houses, aligned with existing research ethics review processes.

1. Introduction

In the context of today's global race to dominate the development and regulation of AI, academic AI research is a high-priority target for state threat actors. The UK Security Service (MI5) describes state threats as "overt or covert actions by foreign governments which fall short of direct armed conflict with the UK but go beyond peaceful diplomacy and expected statecraft to harm or threaten the safety or interests of the UK or our allies."[1] The primary states that currently pose the greatest threat to UK national security are China, Russia, Iran and North Korea.[2]

This section situates academic AI research in the context of state threats and provides an overview of the research methodology used for this project.

1.1 AI as a national asset

The UK Government has signalled its intent to exploit AI as a key tool to grow the economy and improve public services. Maintaining a world-class AI research ecosystem is critical to these ambitions.

AI is one of the five critical technologies the UK Government outlined in the Science and Technology Framework.[3] The Government has previously declared its ambition to become a global leader in AI – an "AI superpower."[4] The creation of state-backed institutes – such as the Advanced Research and Invention Agency and the AI Security Institute (AISI) – and the hosting of the inaugural AI Safety Summit demonstrate an ongoing commitment to this goal. The AI Opportunities Action Plan has delivered a roadmap for exploiting AI to enhance growth and productivity in the UK.[5]

Investment in UK AI R&D and core infrastructure for AI is substantial and growing. DeepMind was founded in the UK before being acquired by Google. In 2024, CoreWeave

[1] MI5, "Countering state threats," https://www.mi5.gov.uk/what-we-do/countering-state-threats.

[2] National Cyber Security Centre, *Annual Review 2024*, https://www.ncsc.gov.uk/files/NCSC_Annual_Review_2024.pdf.

[3] Department for Science, Innovation and Technology, *The UK Science and Technology Framework*, 9 February 2024, https://www.gov.uk/government/publications/uk-science-and-technology-framework/the-uk-science-and-technology-framework#identifying-critical-technologies.

[4] HM Government, *National AI Strategy,* September 2021, https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National_AI_Strategy_-_PDF_version.pdf.

[5] Department for Science, Innovation and Technology, "AI Opportunities Action Plan: terms of reference," 26 July 2024, https://www.gov.uk/government/publications/artificial-intelligence-ai-opportunities-action-plan-terms-of-reference/artificial-intelligence-ai-opportunities-action-plan-terms-of-reference.

committed £1 billion to expand operations in the UK and announced that its new European headquarters would be based in London.[6] OpenAI established its first international office in London.[7] In 2023, Microsoft committed £2.5 billion to expand AI data centre infrastructure in the UK.[8] As of May 2024, the UK accounted for roughly 50% of all private capital investment in AI in Europe.[9]

Investment is also being made in the academic sector, albeit on a smaller scale. In 2023, UKRI allocated £117 million to 12 new AI Centres for Doctoral Training.[10] In early 2024, £80 million was allocated to 9 new Engineering and Physical Sciences Research Council AI research hubs led by UK universities.[11] However, the Government will not be taking forward £800 million of funding for an exascale computer at Edinburgh University.[12] A review into the Future of Compute identified that UK public provision did not meet demand, and the UK requires sustained investment in infrastructure, software and skills.[13]

Nevertheless, the UK is a global leader in AI R&D. Stanford University's AI Index Report lists the UK as a leading source of influential AI models, behind only the US and China.[14] Times Higher Education ranks three UK universities in the global top ten for computer science.[15] Three UK universities also rank in QS World University Rankings' top 20 for data science and artificial intelligence.[16] AISI is the world's first and largest AI safety institute. Government priorities, investment, infrastructure and talent all make the UK an attractive target for states that are interested in acquiring AI capabilities – through legitimate means or otherwise.

[6] HM Treasury, "Leading tech firms invest over £2 billion in the UK in one week," 10 May 2024, https://www.gov.uk/government/news/leading-tech-firms-invest-over-2-billion-in-the-uk-in-one-week.

[7] OpenAI, "Introducing OpenAI London," 28 June 2024, https://openai.com/index/introducing-openai-london/.

[8] HM Treasury, "Boost for UK AI as Microsoft unveils £2.5 billion investment," 30 November 2023, https://www.gov.uk/government/news/boost-for-uk-ai-as-microsoft-unveils-25-billion-investment.

[9] HM Treasury, "Leading tech firms invest over £2 billion in the UK in one week," 10 May 2024, https://www.gov.uk/government/news/leading-tech-firms-invest-over-2-billion-in-the-uk-in-one-week.

[10] UK Research and Innovation, "£100m boost in AI research will propel transformative innovations," 6 February 2024, https://www.ukri.org/news/100m-boost-in-ai-research-will-propel-transformative-innovations/.

[11] Department for Science, Innovation and Technology, "Britain to be made AI match-fit with £118 million skills package," 31 October 2023, https://www.gov.uk/government/news/britain-to-be-made-ai-match-fit-with-118-million-skills-package.

[12] Cliff Saran, "Labour drops Edinburgh exascale supercomputer but funds more AI," *Computer Weekly*, https://www.computerweekly.com/news/366600078/Labour-drops-Edinburgh-exascale-supercomputer-but-funds-more-AI.

[13] Department for Science, Innovation and Technology, *Independent Review of The Future of Compute: Final report and recommendations*, 6 March 2023, "https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts.

[14] Stanford University, "Artificial Intelligence Index Report 2024," https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf.

[15] Times Higher Education, "World University Rankings 2025," https://www.timeshighereducation.com/world-university-rankings/latest/world-ranking#!/length/25/subjects/3081/sort_by/rank/sort_order/asc/cols/stats.

[16] QS, "QS World University Rankings by Subject 2024: Data Science and Artificial Intelligence," https://www.topuniversities.com/university-subject-rankings/data-science-artificial-intelligence.

1.2 International collaboration

International research collaboration is a core strength of the UK's academic ecosystem, as it enables world-leading research, teaching and recruitment. Globally, the proportion of academic publications with international co-authorship has steadily risen in the past decade.[17] International collaboration fosters both economic and social benefits, and can directly impact a university's global ranking.[18] In 2021, 60.4% of the UK's research outputs had an international co-author.[19] According to CSET Georgetown's Country Activity Tracker, China is now the UK's leading partner in co-authorship of AI-related academic publications, having overtaken the US in 2020.[20]

Commentators argue that universities are a source of soft power, as they strengthen economic cooperation and international engagement.[21] Investment in academia can enable high-quality research – which can, in turn, attract foreign collaboration and investment. Singapore, for example, has significantly invested in global R&D in the past 25 years[22] and remains the only Asian country with a university in the top 15 of the QS World University Rankings.[23] The UK Science and Innovation Network operates in Singapore[24] and in 2023 a new strategic partnership was signed between Singapore and the UK to enhance research cooperation in areas such as critical technologies.[25] In early 2024, Imperial College London

[17] Times Higher Education, *International Research Collaboration: Motivators, Enablers and Barriers*, 2023, https://www.timeshighereducation.com/sites/default/files/theconsultancy_intlcollaborationsurvey_report_v5.pdf.
[18] Ibid.
[19] National Protective Security Authority, *Trusted Research Guidance for Academics,* 2024, https://www.npsa.gov.uk/system/files/trusted-research-guidance-for-academia-digital-compressed-1.pdf.
[20] Emerging Technology Observatory, "Country Activity Tracker (CAT): Artificial Intelligence," 18 December 2024, https://cat.eto.tech/.
[21] James Coe, "Turning scrutiny on security into a new security strategy," Wonkhe, 5 September 2024, https://wonkhe.com/blogs/turning-scrutiny-on-security-into-a-new-security-strategy/; Ayu Anastasya Rachman, "Why countries should leverage universities as a new force in global diplomacy," *The Conversation*, 19 September 2020, https://theconversation.com/why-countries-should-leverage-universities-as-a-new-force-in-global-diplomacy-138717.
[22] Rachman, "Why countries should leverage universities as a new force in global diplomacy."
[23] QS, "QS World University Rankings 2024," https://www.topuniversities.com/world-university-rankings/2024.
[24] Foreign, Commonwealth & Development Office, "Singapore: UK Science and Innovation Network summary," 5 February 2025, https://www.gov.uk/government/publications/uk-science-and-innovation-network-country-snapshot-singapore/uk-science-and-innovation-network-country-snapshot-singapore.
[25] Prime Minister's Office, "Joint Declaration by the Prime Ministers of the Republic of Singapore and the United Kingdom of Great Britain and Northern Ireland on a Strategic Partnership," 9 September 2023, https://www.gov.uk/government/publications/uk-singapore-joint-declaration-9-september-2023/joint-declaration-by-the-prime-ministers-of-the-republic-of-singapore-and-the-united-kingdom-of-great-britain-and-northern-ireland-on-a-strategic-part.

opened a research and innovation centre in Singapore.[26] International academic engagement can also provide important insights into other states' innovation capabilities.

1.3 State threats to UK universities

Increasingly, however, academic institutions are faced with the serious threat of exploitation by state threat actors. These actors use espionage, theft and duplicitous collaboration to develop their own technological capabilities and keep pace with R&D in world-leading universities. Warnings have been issued to universities by MI5, the Federal Bureau of Investigation and the European Commission.[27] In an unprecedented joint statement from the heads of Five Eyes[28] countries' intelligence services in October 2023, China was accused of "sustained" and "sophisticated" IP theft.[29]

The 2023 Intelligence and Security Committee of Parliament report on China states that China "directs or steals" academic research to build or shortcut expertise and gain an advantage.[30] In April 2024, MI5 Director-General Ken McCallum gave a briefing to senior Russell Group university staff, warning that research could be targeted for theft by states intending to enhance their own economic and military capabilities.[31] Following this, the then-deputy prime minister launched a consultation on measures to protect UK universities from the national security threats posed by foreign states accessing research with dual-use potential.[32]

UUK states that "threats from hostile state actors are the most prominent security issue universities face."[33] Research, data and expertise can be targeted, transferred or compromised through illegitimate (theft) and legitimate (collaboration and knowledge

[26] Stephen Johns, "Imperial opens first overseas research and innovation centre in Singapore," Imperial University, 8 January 2024, https://www.imperial.ac.uk/news/250533/imperial-opens-first-overseas-research-innovation/.

[27] Nathan Williams, "Foreign states targeting UK universities, MI5 warns," *BBC News*, https://www.bbc.co.uk/news/uk-68902636; MI5, "Joint address by MI5 and FBI Heads," 6 July 2022, https://www.mi5.gov.uk/joint-address-by-mi5-and-fbi-heads; European Commission, "Commission proposes new initiatives to strengthen economic security," 24 January 2024, https://ec.europa.eu/commission/presscorner/detail/en/IP_24_363.

[28] The Five Eyes alliance comprises the UK, the US, Canada, Australia and New Zealand.

[29] Zeba Siddiqui, "Five Eyes intelligence chiefs warn on China's 'theft' of intellectual property," *Reuters*, 18 October 2023, https://www.reuters.com/world/five-eyes-intelligence-chiefs-warn-chinas-theft-intellectual-property-2023-10-18/.

[30] Intelligence and Security Committee of Parliament, *Intelligence and Security Committee of Parliament: China*, 13 July 2023, https://isc.independent.gov.uk/wp-content/uploads/2023/07/ISC-China.pdf.

[31] Williams, "Foreign states targeting UK universities, MI5 warns."

[32] Cabinet Office, "Government to launch new consultation to protect UK universities from security threats," 26 April 2024, https://www.gov.uk/government/news/government-to-launch-new-consultation-to-protect-uk-universities-from-security-threats.

[33] Universities UK, "Security and risk: how universities can protect their research and people," 8 June 2023, https://www.universitiesuk.ac.uk/topics/funding-finance-and-operations/security-and-risk-how-universities-can.

acquisition) means. State actors may attempt to access researchers and their work through traditional engagement such as research placements or conferences.[34] Institutions may target academics to undertake strategically beneficial research.[35] Dual-use research[36] and applied research that may have sensitive applications are particularly at risk of misappropriation and could be used to undermine UK national security or to target or harm other populations.[37]

Example of a state threat

In December 2020, a UK newspaper reported that a laboratory at a UK University – part-funded by the UK Ministry of Defence – working on quantum technology had nurtured a link with the Chinese military university – the National University of Defence Technology (NUDT). The article claimed that a Chinese national had "unrivalled access to the lab's most sensitive projects" and that the "NUDT had paid for his studies in the UK."[38]

Researchers and institutions that collaborate with hostile states (knowingly or unknowingly) face reputational risks, loss of IP and constraints on – or interference with – their academic freedom.[39] If research is stolen and exploited by state actors for military or unethical purposes, this may make it more difficult for academics and institutions to secure funding or sponsorship, or to access sensitive data.[40]

1.4 Research aims and methodology

A large body of subject-agnostic guidance on Trusted Research has been developed by the NPSA.[41] This CETaS Research Report seeks to complement such existing work by producing tailored guidance for academics working on AI research.

The main research aims were to:

1. Identify the unique vulnerabilities of academic AI research in the UK.

[34] National Protective Security Authority, *Trusted Research Guidance for Academics*.
[35] Ibid.
[36] Dual-use refers to technology that can be applied for both civilian and military purposes.
[37] National Protective Security Authority, "State Threats in Academia," 22 January 2024, https://www.npsa.gov.uk/state-threats-academia.
[38] Ibid.
[39] University of Oxford, "Trusted Research," https://researchsupport.admin.ox.ac.uk/trusted-research#tab-2697401.
[40] National Protective Security Authority, *Trusted Research Guidance for Academics*.
[41] See Annex A for a summary of existing guidance.

2. Explore the national security risks associated with the theft and acquisition of, and interference with, academic AI research.
3. Make recommendations for strengthening the resilience of the UK's AI research ecosystem to state threats.

This study was conducted over a 7-month period from May 2024 – November 2024. Data collection involved the following core research activities:

- **A systematic literature review** of academic and grey literature to establish the context (including the threat landscape and existing guidance on research security).
- **Interviews** with government and academic representatives to understand perspectives on the threat landscape.
- **A focus group** with academic representatives and professional services staff to gather perspectives on the barriers to research security.
- **Scenario development** to illustrate the threat landscape for AI researchers.

The following themes are out of scope for this research:

- AI research conducted by security cleared researchers in secure working areas.
- AI research conducted by commercial (or industry) bodies or university spin-out companies.

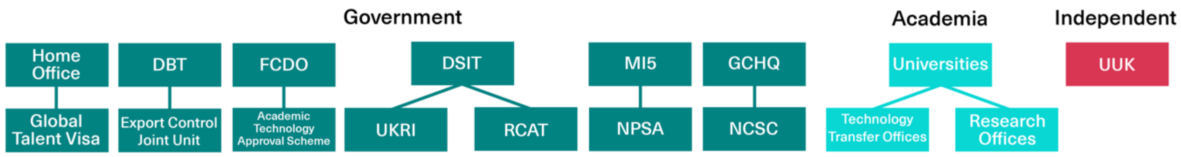
The remainder of the report is structured as follows. Section 2 of the report outlines the current research security ecosystem. Section 3 discusses the risks and vulnerabilities applicable to AI research, using fictitious scenarios to illustrate the threat. Section 4 outlines barriers and enablers to building resilience to the threat within the academic community. Section 5 considers best practice case studies from other industries and countries, and presents a maturity model designed to support academic institutions that conduct AI research. Finally, Section 6 provides a series of policy recommendations for strengthening the resilience of the UK's AI research ecosystem.

Annex A summarises current research security measures and existing guidance.

2. The UK Research Security Ecosystem

This section describes the UK's research security ecosystem, summarises existing legislation and guidance, and presents research findings on the systemic barriers to resilience building.

Figure 1. Institutions involved in research security



Source: Authors' analysis.

As shown in Figure 1, the UK research security ecosystem is a complex web of government, academic and independent bodies. The large number of government departments involved in research security is a barrier to cohesion across policy teams and creates friction for academics and professional services staff seeking guidance and support.[42] A key challenge for researchers is that, despite this complex ecosystem, risk assessment is often left to individual academics who do not necessarily have knowledge of the threat landscape.[43] It is difficult for researchers to foresee and quantify the risks stemming from early-stage research – and understanding how research may be exploited by adversaries is not an easy task.

Due diligence is a critical enabler of research security for AI researchers. However, due diligence for academic research security can be more resource-intensive than other forms of due diligence due to the myriad considerations required to understand the potential risks.[44] While a body of guidance exists to support academics, it is not easily accessible in one place. Research participants also expressed frustration at the large administrative burden that comes with due diligence, which can feel restrictive to academics and professional services staff.[45]

[42] CETaS Workshop, 1 November 2024; author interview with academic participant, 7 November 2024.
[43] CETaS Workshop, 1 November 2024; author interview with academic participant, 7 November 2024.
[44] Marwaha et al., *Complex Collaborations*, Association of Research Managers and Administrators, March 2023, https://arma.ac.uk/wp-content/uploads/2023/03/Trusted-Report_Booklet_v7.pdf.
[45] CETaS Workshop, 1 November 2024; author interview with academic participant, 7 November 2024.

More specific guidance for operationalising risk management related to research security would be useful for AI researchers.[46] New, tailored guidance should be developed by the AI research community (with input from the Government) to ensure it remains resilient to the dynamic threat and can be adapted and tailored to the institutional context.[47] A shared due diligence repository would also be a useful knowledge-sharing and support tool for the research community.[48]

Continuing to invest in awareness-raising initiatives and support mechanisms for academia should be a key priority for Government, to provide AI researchers with easily accessible and clear information on the threat landscape.

Best Practice Case Study: Research Collaboration Advice Team

The RCAT describes its work as "a collaboration between the government and academia providing research institutions with a first point of contact for official advice about national security risks linked to international research."[49]

Based within DSIT, the RCAT is an advisory body that helps universities confidently collaborate with international partners while safeguarding their research. The RCAT provides a first point of contact for research leaders seeking advice and information, and maintains the principles of academic freedom and independence wherever possible, while supporting the identification of national security concerns and appropriate mitigations.

Research participants praised the RCAT as an important support function and key link between the Government and academia. However, they raised concerns about the team's capacity to provide guidance as the academic community develops a greater awareness of the risks.

[46] Marwaha et al., *Complex Collaborations*.

[47] CETaS Workshop, 1 November 2024; Marwaha et al., *Complex Collaborations*.

[48] Marwaha et al., *Complex Collaborations*; CETaS Workshop, 1 November 2024.

[49] Research Collaboration Advice Team, "About us," https://www.gov.uk/government/organisations/research-collaboration-advice-team/about.

2.1 Existing legislation

The legislative framework that governs academic research in the UK is built on a fragmented legal landscape and was not designed with state threats in mind.[50] This means that a broad range of legislative provisions are applicable to academic AI research. Non-legislative official guidance on research security has been developed to support researchers. (See Annex A for a summary of existing legislation and guidance.)

The broad nature of existing legislation and guidance respects academic freedom – a key principle of the UK's higher education sector that allows researchers to conduct their work without interference or punishment.[51] It also enables a flexible and principles-based approach to research security. This is beneficial because technological developments and dynamic threats will always outpace risk-based regulation.

Throughout the 2010s, researchers were encouraged to collaborate internationally and partnerships between Chinese organisations and UK researchers were considered part of a "Golden Era" in strategic cooperation.[52] In this context, it is important that the legislative environment now remains stable, so institutions and individual researchers are given time to adapt and acclimatise to updated guidance.

Compliance with legislation is enforced through prosecution and mandatory notification obligations, but there is no inspection or monitoring regime for voluntary guidance such as that issued by the NPSA. Terms and conditions for UKRI funding is another mechanism utilised for Trusted Research guidance compliance, though a lack of specificity weakens this mechanism. The roles of devolved regulators (Office for Students, Commission for Tertiary Education and Research, Scottish Funding Council, Higher Education Division) in compliance are unclear.[53] At present, there is a lack of incentives for researchers to follow non-binding government-issued guidance.

[50] Ada Lovelace Institute, *Regulating AI in the UK*, July 2023, https://www.adalovelaceinstitute.org/wp-content/uploads/2023/09/ALI_Regulating-AI-in-the-UK_2023.pdf.

[51] University and College Union Scotland, "Charter on Academic Freedom," https://www.ucu.org.uk/media/4160/Scottish-Charter-on-Academic-Freedom/pdf/Academic_Freedom_charter.pdf.

[52] HM Treasury, "Economic talks herald Golden Era in UK-China relations," 16 December 2017, https://www.gov.uk/government/news/economic-talks-herald-golden-era-in-uk-china-relations.

[53] Graeme Atherton, Joe Lewis and Paul Bolton, *Higher education in the UK: Systems, policy approaches, and challenges*, House of Commons Library, 15 July 2024, https://researchbriefings.files.parliament.uk/documents/CBP-9640/CBP-9640.pdf.

3. Risks to AI Research

This section outlines the value of AI research to state threat actors and considers characteristics that may be targeted for theft and acquisition, sabotage, and interference. Fictitious scenarios designed to illustrate the threat are also introduced.

3.1 The value of AI

AI technologies and associated datasets are high-value targets for states seeking to replicate or undermine a nation's capabilities. The theft or acquisition of AI models could allow adversaries to gain capabilities quickly and cheaply. Beyond considerations of national security is the ethical responsibility to ensure that UK-led research is not being used by states to harm their own populations or others.

AI is a dual-use technology with applications across a huge number of fields from misinformation and surveillance to code generation and precision medicine.[54] By reverse-engineering models, adversaries can reproduce or improve upon proprietary systems, potentially neutralising a technological advantage. For example, tools for countering malicious uses of AI systems (e.g. synthetic media detection) could be reversed-engineered or subverted, allowing adversaries to evade detection more effectively. Accessing models might also enable state actors to exploit AI-driven systems, evading detection or bypassing defences with tailored cyberattacks.

The underlying sensitive datasets used to train AI models (e.g. biometric data or satellite imagery) could provide strategic insights that impact defence planning and intelligence efforts. Personal data can also be extracted to undermine individuals' privacy. For example, several commentators have raised privacy concerns related to data aggregation for smart cities.[55] All data, not just sensitive data, are valuable and have the potential to cause harm.

[54] Alexei Grinbaum and Laurynas Adomaitis, "Dual use concerns of generative AI and large language models," *Journal of Responsible Innovation* 11, no.1 (2024), https://www.tandfonline.com/doi/full/10.1080/23299460.2024.2304381; Ardi Janjeva, Anna Gausen, Sarah Mercer and Tvesha Sippy, "Evaluating Malicious Generative AI Capabilities: Understanding inflection points in risk," *CETaS Briefing Papers* (July 2024).

[55] Ash Johnson, "Balancing Privacy and Innovation in Smart Cities and Communities," Information Technology & Innovation Foundation, 6 March 2023, https://itif.org/publications/2023/03/06/balancing-privacy-and-innovation-in-smart-cities-and-communities/; Nate Lavoy, "Addressing the National Security Risks of Bulk Data in the Age of AI," RAND, 23 August 2024, https://www.rand.org/pubs/commentary/2024/08/addressing-the-national-security-risks-of-bulk-data.html; Arvind Narayanan and Vitaly Shmatikov, "How To Break Anonymity of the Netflix Prize Dataset," *arXiv*, 22 November 2007, https://arxiv.org/abs/cs/0610105.

Data breaches at a DNA testing firm,[56] a breast cancer clinic[57] and an 'AI Girlfriend' service[58] all demonstrate that data, once collected, is always a potential vector for harm.[59] Good cybersecurity hygiene can mitigate the impact of data breaches, as can building awareness of how such data enables hostile actors.[60]

3.2 Data as a vulnerability

A key constraint for AI researchers is access to high-quality data. Published open-source datasets vary in size, exhaustivity (capturing an entire population), relationality (features allowing for integration with other datasets) and velocity (the speed of collection and use). They often lack metadata relating to provenance – how data was obtained or collected.[61] There are no common standards or quality assurance measures in place to evaluate open-source data. Without fully understanding the data collection process, researchers cannot be fully informed of any biases that may reside within the data set. Open-source datasets can also be the target of malicious attacks. Recent research from DeepMind and NVIDIA has demonstrated the risk of 'frontrunning poisoning' attacks, which target web-scale datasets that periodically snapshot crowdsourced content (e.g. Wikipedia) and inject malicious content just before scrapes are conducted.[62]

In the absence of good quality and reliable training data, researchers are increasingly turning to synthetic data to top up or fine-tune existing models. Overreliance on synthetic data, however, risks overfitting to characteristics in the synthetic data generation process. Large language models (LLMs) may also be used to label text datasets, reportedly to the same or higher standards than that of human labellers.[63] In this instance, the datasets would likely be subject to bias and hallucinations.

[56] Mack DeGeurin, "Hackers got nearly 7 million people's data from 23andMe. The firm blamed users in 'very dumb' move," *The Guardian*, https://www.theguardian.com/technology/2024/feb/15/23andme-hack-data-genetic-data-selling-response.

[57] Marianne Kolbasuk McGee, "Breast Cancer Patients Sue Over Breached Exam Photos, Data," CIO, 14 March 2023, https://www.cio.inc/breast-cancer-patients-sue-over-breached-exam-photos-data-a-21431.

[58] Pieter Arntz, "AI girlfriend site breached, user fantasies stolen," MalwareBytes Labs, 9 October 2024, https://www.malwarebytes.com/blog/news/2024/10/ai-girlfriend-site-breached-user-fantasies-stolen.

[59] Alaina Demopoulos, "There are no serious safeguards': can 23andMe be trusted with our DNA?," *The Guardian*, https://www.theguardian.com/technology/2024/feb/17/23andme-dna-data-security-finance.

[60] National Cyber Security Centre, "Advice & guidance," https://www.ncsc.gov.uk/section/advice-guidance/all-topics.

[61] Jack Hughes, Yi Ting Chua and Alice Hutchings, "Too Much Data? Opportunities and Challenges of Large Datasets and Cybercrime," 2021, https://www.cl.cam.ac.uk/~ah793/papers/2021_too_much_data.pdf.

[62] Carlini et al., "Poisoning Web-Scale Training Datasets is Practical," *arXiv*, 6 May 2024, https://arxiv.org/pdf/2302.10149.

[63] Refuel, "LLMs can structure data as well as humans, but 100x faster," 14 June 2023, https://www.refuel.ai/blog-posts/llm-labeling-technical-report.

Existing data protection practices provide a good starting point for addressing the above vulnerabilities, but researchers must explore data considerations across the whole data lifecycle – from acquisition to disposal. Adopting strict revision control mechanisms and engaging with a security expert will embed good practice in a research project from the start. The proposed National Data Library could also be an asset to academic AI researchers seeking to access high-quality, reliable datasets for training AI models.[64]

3.3 Risks of AI advancement

The rapid advancement of AI research means that tools, frameworks and best practices are constantly evolving, often faster than in equivalent engineering fields. This can lead to gaps in reliability and robustness because – unlike traditional engineering, where mature tools allow for informed trade-offs – AI researchers are sometimes limited to using the latest or only available tool, which may not have been fully tested. The evolving nature of AI research also means that engineering practices and standards are still catching up. This lack of standardisation can result in inconsistent outcomes, making it challenging to ensure reliability.

AI research presents unique risks beyond traditional software development and deployment environment risks:

- **Impact.** The complex data dependencies and adaptive nature of machine learning (ML) algorithms mean that even subtle attacks – like data poisoning or adversarial manipulation – can lead to significant, often undetected impacts on model performance and reliability.
- **Security.** AI workflows often rely on specialised libraries, less-mature compilers and high-performance hardware that may lack the stability and maturity of standard software environments. The risks of working with less mature components include a lack of built-in security measures and variable responses to discovered vulnerabilities. In addition, AI pipelines require custom data handling and storage solutions for managing very large datasets.
- **Dynamic nature of AI.** The dynamic nature of some AI systems introduces risks like model drift, edge cases and safety issues, all of which complicate the processes of verification and validation compared to traditional software systems.

[64] Department for Science, Innovation and Technology, "AI Opportunities Action Plan: government response," 13 January 2025, https://www.gov.uk/government/publications/ai-opportunities-action-plan-government-response/ai-opportunities-action-plan-government-response.

The table below outlines several attack types relevant to ML models.[65]

Attack	Description	Purpose
Data poisoning	Adversaries manipulate or introduce false data into training datasets, causing models to behave incorrectly or unexpectedly.	To erode accuracy or confidence, or to introduce a bias or weakness that can later be exploited.
Model inference	Adversaries reverse-engineer the model to recover private or sensitive data, or to understand its decision-making processes.	To extract information about the model and/or its training data. Model extraction can also be used to repeatedly query a model, use the answers to label a dataset and subsequently train a surrogate version of the model.
Adversarial input	Small, carefully created changes to inputs used to confuse a model, making it misclassify or misinterpret the input.	To trick an AI system into misidentifying a threat. Unlike data poisoning, this does not involve manipulating training data but requires knowledge of how the model works.
Theft	Analysis, replication or modification of a model. The model could later be reverse-engineered to recover sensitive or personal data, or to gain strategic insights into its behaviour.	Insider threat or remote attack could allow an adversary to steal training data or a compiled model.

[65] For further examples and information, please see the following sources: “Navigate threats to AI systems through real-world insights,” MITRE ATLAS, https://atlas.mitre.org; OWASP, “OWASP Machine Learning Security Top Ten,” https://owasp.org/www-project-machine-learning-security-top-10/; OWASP, “OWASP Top 10 for Large Language Model Applications,” https://owasp.org/www-project-top-10-for-large-language-model-applications/.

The implementation of proactive security measures throughout the research lifecycle is essential – strong access controls and data encryption should be considered at every stage. Rigorous logging and monitoring systems should be established to detect anomalies and unauthorised access early. Judgements about the security and reliability of subcomponents (and their vendors) should be considered when setting up a research environment. Collaboration with security experts during the early phases of research will also help embed security best practice.

3.4 Testing and verification

Testing and verifying to ensure model safety and guard against tampering is important yet challenging for AI researchers. ML systems are often designed for situations where the definition of 'correct' is elusive – it is difficult to determine a fixed and expected outcome for every input. Probabilistic behaviour will lead to a model generating different outputs for the same input. The output will depend on the model's training, the underlying data distribution and random initialisation factors. This means that testing centres on assessing the reliability of the model's behaviour under diverse conditions rather than verifying specific rules. As ML models are trained on historical data, their performance and behaviours are highly dependent on the quality, completeness and relevance of that data.

Many powerful ML models (especially deep learning models) operate as black boxes.[66] This opacity makes it difficult to trace reasoning and complicates debugging and validation. These factors make it more challenging to ensure model reliability and safety, particularly in critical applications, and make it trickier to detect whether an adversary has tampered with a model – or dependent datasets.

A range of approaches can address the above challenges:

- **Data validation** will help pre-empt post-deployment issues of susceptibility to edge cases and bias.
- **Stress testing** exposes models to edge cases and ensures resilience under atypical conditions.

[66] The 'black box problem' refers to an opaque system where calculation processes are not interpretable to the user.

- **Fairness assessments**[67] detect and mitigate bias by analysing model outputs across demographic groups.
- **Explainability tools**[68] like SHAP and LIME may help provide transparency in model decisions.
- **Mechanistic interpretability**[69] could provide insights into a model's inner workings, and continuous monitoring can help detect model drift.

The rapid progression and development of AI research requires a cautious and security-aware approach. Access to quality data, the proactive use of security measures, and continuous monitoring of model behaviour can help mitigate the outlined risks.

Further research could also be undertaken by the AI research community to explore topics related to securing AI and protecting IP such as risk management frameworks or approaches for quantifying or detecting whether an ML model has been subject to tampering.

3.5 Fictitious threat scenarios

The following fictitious scenarios were developed by the research team using insights from research participants to illustrate the current threat landscape. They are not intended to be representative of any existing or historic real-world threat scenarios.

[67] Arpit Narain, "Unmasking Bias – Assessing Fairness in Large Language Model," Medium, 7 June 2023, https://medium.com/@arpitnarain/unmasking-bias-assessing-fairness-in-large-language-models-a722624e4483.

[68] Marco Tulio Ribiero, Sameer Singh and Carlos Guestrin, "Why Should I Trust you?, Explaining the Predictions of Any Classifier," *arXiv*, 16 February 2016, https://arxiv.org/abs/1602.04938; Scott Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv*, 22 May 2017, https://arxiv.org/abs/1705.07874.

[69] Neel Nanda, "A Comprehensive Mechanistic Interpretability Explainer & Glossary," https://www.neelnanda.io/mechanistic-interpretability/glossary.

Scenario 1: Repurposed AI research

Professor Y is a senior academic at University A, researching AI-powered text recognition. Their next project involves developing an ML model to identify and replicate unique handwriting for screen reader software. Professor Y has struggled to obtain grant funding for the project, but discovers a funding opportunity with Company A. Company A is a multinational technology company headquartered outside the UK with an excellent reputation in ML research. Company A publicly advocates for open research, and a prerequisite of the funding opportunity is that any training data and source code must be stored in a public repository on GitHub. Professor Y accepts the funding opportunity and shares the research on GitHub.

A year later, Professor Y reads an online news article about a large-scale targeted attack on UK Government staff who have had their personal details stolen and sold on the dark web. Reports state it is likely that a sophisticated AI tool was used to forge their signatures to access personal documents.

Scenario 2: Corporate espionage

Student X concludes a PhD programme and accepts a visiting researcher position at a prestigious AI research institute alongside a salaried role at their former university. They spend two years developing their research and engaging widely across the institute. Following a networking event, Student X receives an email inviting them to apply for a remote research role at a newly established start-up looking to commission research on the UK's AI ecosystem. Payment is issued on the delivery of each project for maximum flexibility.

Student X responds positively and is informed that no interview is required, given the strength of their PhD, their status as a visiting researcher and their postdoctoral role. Student X's first project is an open-source mapping exercise on research across the institute, including an organogram of key funders, senior management and research leads. Student X receives generous payment following project delivery. For the second project, the start-up is keen to understand more about Student X's network for a market research contract it has recently obtained. For example, which members of the institute dictate research requirements and strategy? Student X accepts the tasking, and payment follows.

For the third project, Student X is asked to compile an attendee list for a workshop at a leading commercial AI conference. Student X is asked to provide names of colleagues that interact widely across the Government and known government contacts. Student X is uncomfortable with compiling the attendee list and searches for advice from the institute. There is no dedicated internal advisory role on research security, so Student X asks for advice at a visiting researcher roundtable. Several participants are enthusiastic about increased engagement with commercial and government entities. Student X sends the attendee list to the start-up. Several weeks pass without response, and Student X follows up over email – an automated message is returned stating the recipient address is no longer valid.

Scenario 3: Sabotage

Visitor A is a visiting researcher temporarily attached to a renowned Biometrics Research Centre to work on Project Y. Visitor A is friendly and keen to develop relationships with staff across the department. Researcher B is lead researcher on Project X, a new facial recognition system being developed at the Centre. Visitor A offers informal assistance to design and implement novel evaluation techniques for Project X. Researcher B allows Visitor A temporary access to the private GitHub repository, so they can contribute to the main development branch.

Visitor A, using knowledge of the structure and parameters of the facial recognition system, leverages a generative adversarial framework to conduct an attack. Using techniques like those demonstrated in physical-world attacks on neural networks, Visitor A applies subtle imperceptible perturbations to a set of facial images. Over several iterations, the generative adversarial framework refines these changes until the system consistently misclassifies these faces as other individuals.

Researcher B never discovers that Visitor A was being paid by a state-sponsored group to interfere with facial recognition research. Temporary access to the GitHub repository allows individuals associated with Visitor A's handlers to bypass the new facial recognition system, intended to be rolled out at UK airports.

4. Building Resilience

This section presents findings on the current barriers to research security faced by AI researchers and outlines approaches to building resilience in the academic community.

4.1 Academic culture

A fundamental tension exists between research security and the publication of high-quality research. The 'publish or perish' aphorism denotes the professional pressure that academics face to publish their work. This pressure can come from several sources. Reproducible research is a key principle underpinning the scientific method; it requires the publishing of the original data and methods used by a researcher. Editorial boards of academic journals will often reject submissions where data and code are not made available.[70] Successful publications draw talent and funding to institutions, and citations and the impact of publications are commonly considered key performance indicators.[71]

AI researchers, therefore, are encouraged to make their research publicly available and disincentivised from withholding research. While reproducible research can support collaboration and transparency, this practice embeds an inherent vulnerability in academic culture. One short-term mitigation is standardising the practice of conducting pre-publication risk assessments in areas of special concern, such as AI research.[72] Scrutiny committees (as in the below case study) could function similarly to ethics committees and, with early involvement in the research process, could help researchers identify and mitigate risks.

Best Practice Case Study: Scrutiny Committee

Imperial College London has established a scrutiny committee to provide support in reviewing "interactions or activities deemed to be sensitive." The committee will review

[70] CETaS Workshop, 1 November 2024.
[71] CETaS Workshop, 1 November 2024.
[72] Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *arXiv*, February 2018, https://arxiv.org/pdf/1802.07228.

evidence and advise on “conditions for continuation, suspension, or even termination of the proposed activity.”[73]

In the longer term, a cultural shift is required so that research security is perceived by AI researchers as enabling and essential for high-quality research. Both government and professional services staff have previously found it difficult to articulate the relevance of research security to academics.[74] Compounding this issue, peer-to-peer academic collaborations often take place informally – with no funding or support from professional services teams – due to the information-sharing culture of academia and early-stage research.[75] A stronger, well-informed self-regulation approach within the academic community – built on increased risk awareness – will create incentives to follow existing guidance and strike the right balance between academic freedom and research security.[76] This will require a culture change in the academic community and input from the Government to mitigate bias towards publishing.

A ‘whole system’ approach to identify and involve all relevant bodies in discussions on AI and research security would mitigate any disconnect between researchers and government actors. The Government should build relationships with publishing houses and editorial boards of relevant data science and AI journals to discuss solutions to the reproducible research problem and potential vulnerabilities outside the control of individual researchers (such as the peer review process). Institutions should ensure they have clear policies on due diligence and risk management that build a culture of confident collaboration and collective responsibility.[77] Responsibility for raising awareness of research security among more junior staff and students at individual institutions should sit with the senior leadership at the institutional and departmental levels.[78]

[73] Universities UK, “*Managing risks in international research and innovation: An overview of higher education sector guidance*,” June 2022, https://www.universitiesuk.ac.uk/sites/default/files/field/downloads/2022-06/managing-risks-in-international-research-and-innovation-uuk-cpni-ukri_1.pdf.

[74] Author interview with government participant, 30 July 2024; Marwaha et al., *Complex Collaborations*; CETaS Workshop, 1 November 2024.

[75] Marwaha et al., *Complex Collaborations*; author interview with government participant, 1 November 2024; CETaS Workshop, 1 November 2024.

[76] Alexei Grinbaum and Laurynas Adomaitis, “Dual use concerns of generative AI and large language models,” *Journal of Responsible Innovation* 11, no.1 (2024), https://www.tandfonline.com/doi/full/10.1080/23299460.2024.2304381.

[77] Universities UK, *Managing Risks in Internationalisation: Security Related Issues*, October 2020, https://www.universitiesuk.ac.uk/sites/default/files/field/downloads/2021-07/managing-risks-in-internationalisation.pdf.

[78] Author interview with government participant, 1 November 2024.

4.2 Risk perceptions

Culture change will also involve raising awareness of the risk within institutions. Individual institutions should create space for staff and students to express concerns and discuss risks openly to build a culture of support, awareness and resilience. 'Security champions' within faculties could provide peer-to-peer support and information on the day-to-day applicability of research security risks and guidance. [79] The creation of knowledge-sharing spaces where research-intensive universities can share experiences and provide advice would also be beneficial to the sector.[80] UUK could provide a neutral platform for roundtables on research security. The Government should also signal key research areas where collaboration and research engagement are considered beneficial to the UK and the wider sector. This will help avoid a chilling effect whereby researchers choose not to engage in collaboration due to a lack of understanding or misunderstanding of the risks.

Academics in the UK must also be aware of cultural differences and political sensitivities globally and how they might impact research security or present personal risks to their academic collaborators. For example, digital surveillance has been used by the Chinese state to monitor overseas students.[81] Scholars at Risk has reported a number of incidents of 'attacks' on academic freedom, such as the arrest of academics and students who criticised the Nicaraguan Government, the arrest of academics in Libya for union activities and the dismissal of academics in the US due to their teaching on "disfavoured topics."[82]

Common expectations and norms in the UK – such as adherence to the reproducible research principle – can also make UK AI research more vulnerable to state threats.[83] The Academic Freedom Index uses indicators such as 'institutional autonomy' and 'freedom of academic and cultural expression' to assess levels of academic freedom in 179 countries, and is a useful tool for academics conducting due diligence research.[84]

[79] Universities UK, "Security and risk: how universities can protect their research and people," 8 June 2023, https://www.universitiesuk.ac.uk/topics/funding-finance-and-operations/security-and-risk-how-universities-can.

[80] Author interview with academic participant, 7 November 2024.

[81] Amnesty International, "China: Overseas students face harassment and surveillance in campaign of transnational repression," 13 May 2024, https://www.amnesty.org/en/latest/news/2024/05/china-overseas-students-face-harassment-and-surveillance-in-campaign-of-transnational-repression/.

[82] Scholars at Risk Network, "Free to Think," 2024, https://www.scholarsatrisk.org/resources/free-to-think-2024/.

[83] Author interview with government participant, 1 November 2024.

[84] Katrin Kinzelbach, Staffan I. Lindberg and Lars Lott, *Academic Freedom Index: Update 2024,* Friedrich-Alexander-Universität Institute of Political Science, 2024, https://academic-freedom-index.net/research/Academic_Freedom_Index_Update_2024.pdf.

4.3 Funding and talent retention

A lack of access to funding and poor talent retention also introduce new vulnerabilities relating to research security – by, for instance, creating incentive structures for academics to accept funding from dubious sources or accept better-paid roles at organisations that can then exploit their insight and expertise.

A number of publications have previously discussed the UK higher education funding model in detail.[85] In England, it is estimated that universities supplement the cost of educating undergraduates by £2,500 per student per year.[86] UK universities spend more on delivering research than they receive in funding – and must, therefore, recuperate their costs from sources such as commercial income and international student fees.[87]

Given this context, refusing funding can be perceived as professionally damaging for a researcher. For an individual academic, turning down a funding opportunity could be synonymous with turning down a new research facility, a permanent contract or even a promotion – and thus is potentially career limiting.[88] To compound this issue, the digital skills gap in the UK[89] means the talent pipeline for AI research is poor; this problem is exacerbated as high salaries and research opportunities offered by the private sector limit talent retention within universities.[90]

The importance of a strong talent pipeline directly relates to the porous nature of AI research. Building resilience will not eliminate the threat, but it will create friction for state actors looking to steal or acquire research. Mitigating the impact of research theft and acquisition will require a strong academic sector that can nurture talent and maintain a steady flow of high-quality research on the frontier of technological development.

[85] For further reading on the topic of UK HE funding, please see the following links: https://www.ukri.org/publications/research-financial-sustainability-data/research-financial-sustainability-issues-paper/; https://www.universitiesuk.ac.uk/what-we-do/policy-and-research/publications/financial-sustainability-uk-universities; https://ifs.org.uk/publications/higher-education-finances-how-have-they-fared-and-what-options-will-incoming; https://wonkhe.com/blogs/uuk-has-a-plan-to-fix-research-funding/; https://www.nature.com/articles/d41586-024-03079-w; https://commonslibrary.parliament.uk/higher-education-funding-trends-and-challenges/; https://russellgroup.ac.uk/media/6145/university-business-model-explainer.pdf.

[86] Russell Group, "University business model explainer," 31 August 2023, https://www.russellgroup.ac.uk/policy/policy-briefings/university-business-model-explainer.

[87] Universities UK, "Funding," https://www.universitiesuk.ac.uk/what-we-do/policy-and-research/publications/features/research-and-innovation-facts-and/funding.

[88] Author interview with academic participant, 12 July 2024; CETaS Workshop, 1 November 2024.

[89] King's Trust, "The Prince's Trust: Decoding the Digital Skills Gap," 1 August 2024, https://www.kingstrust.org.uk/about-us/news-views/decoding-digital-skills-gap-report.

[90] Author interview with academic participant, 7 November 2024.

Case Study: The National Quantum Computing Centre

The National Quantum Computing Centre (NQCC) is a research institution funded through UKRI, dedicated to accelerating the development of quantum computing through the coordination and delivery of a technical programme, alongside the commissioning and operation of new facilities.[91] The NQCC aims to address the key engineering challenges involved with scaling quantum computers and engaging and cooperating with industry, the Government and the research community.

Like AI, quantum computing is an inherently dual-use technology. Both fields require collaboration with a wide range of actors and investors due to funding and technical requirements. The NQCC is a trusted authority that maintains a supportive environment for researchers and supports skills development to boost the talent pipeline.

Mitigating the threats to academic AI research – theft, acquisition, sabotage and interference – from state actors will require commitment from academic institutions and strong relationship-building and support from the Government. Short-term mitigations to create friction for state threat actors are needed, alongside investment in a long-term strategy to build the resilience of the academic sector.

[91] National Quantum Computing Centre, "About us," https://www.nqcc.ac.uk/about-us/.

5. Best Practice

This section describes the challenges of safeguarding dual-use technologies, and uses best practice case studies to highlight examples of resilience-building from other sectors and internationally. This section also presents a Maturity Model designed to support academic institutions conducting AI research.

5.1 Managing the dual-use problem

Dual-use items are goods, software and technology that can be used for both civil and military purposes. This definition covers a wide range of research areas, from micro-organisms and toxins to aerospace and propulsion.[92] The dual-use problem is an ethical dilemma which arises when researchers conduct research that may result in knowledge being used in undesirable and unpredictable ways.

Engineering Biology as a dual-use technology

Innovations in gene editing and synthesis (partially attributed to the development of AI) hold immense promise for transforming sectors such as medicine and agriculture.[93] However, the dual-use nature of engineering biology could lead to unprecedented dangers, such as the development of biological weapons.[94] Researchers showed that an AI model designed for drug discovery could be repurposed to seek out toxic substances such as nerve agents.[95]

[92] Export Control Joint Unit, "Export controls: dual-use items, software and technology, goods for torture and radioactive sources," 19 July 2023, https://www.gov.uk/guidance/export-controls-dual-use-items-software-and-technology-goods-for-torture-and-radioactive-sources.

[93] Eric Schmidtapril, "TIME: We Need to be Ready for Biotech's ChatGPT Moment," BioHealth Innovation, 16 April 2024, https://www.biohealthinnovation.org/biohealth-news/17-news/14945-time-we-need-to-be-ready-for-biotech-s-chatgpt-moment; Stewart Patrick and Josie Barton, "Mitigating Risks from Gene Editing and Synthetic Biology: Global Governance Priorities," Carnegie Endowment for International Peace, 16 October 2024, https://carnegieendowment.org/research/2024/10/mitigating-risks-from-gene-editing-and-synthetic-biology-global-governance-priorities.

[94] Roger Brent, T. Greg McKelvey, Jr. and Jason Matheny, "The New Bioweapons," *Foreign Affairs*, 20 August 2024, https://www.foreignaffairs.com/world/new-bioweapons-covid-biology.

[95] Thomas Douglas, "The dual-use problem, scientific isolationism and the division of moral labour," *Monash bioethics review* 32, no.1 (2014), https://pmc.ncbi.nlm.nih.gov/articles/PMC4210745/.

5.1.1 Promoting transparency: Cyber surveillance

While cyber technologies have become indispensable to everyday life, the proliferation of specific technologies poses a significant risk of abuses of human rights, such as the right to privacy. There have been several instances of privacy violations stemming from the sale and distribution of cyber surveillance tools – such as those during the 'Arab Spring,' in which commercial surveillance technologies were used in a discriminatory manner to monitor and identify political opponents and activists.[96] With the goal of strengthening security and transparency, technical guidance was published by the UK Government in 2019 with advice for exporters regarding items that may be used to "intercept communications." This guidance was updated in 2023 to reflect sanctions that banned the export of such items to Russia, Belarus, Myanmar, Iran, Syria and Venezuela.[97] This additional guidance acts as a clear steer from the Government on technologies of concern. As with AI, cyber technologies continue to emerge at an unprecedented rate, so it is essential that controls to tackle proliferation remain transparent, relevant and effective.

5.1.2 Safeguarding: Nuclear technology

It is important to understand how other high-risk sectors have developed preventive measures to combat risk. Nuclear technology presents a unique dual-use circumstance given its history as a military technology with civilian use cases rather than vice-versa.[98] Despite this unique characteristic, some techniques used to safeguard the technology offer interesting lessons for AI research.

Many incidents have shaped nuclear threat perceptions and how the nuclear sector is regulated internationally. One example is A.Q. Khan's nuclear proliferation ring. Khan, who founded Pakistan's nuclear weapons programme, became familiar with European nuclear technologies through multiple visits to enrichment facilities. He left for Pakistan with blueprints for several components, as well as contact details for private suppliers.[99] Khan

[96] "How BAE sold cyber-surveillance tools to Arab states," *BBC News*, 15 June 2017, https://www.bbc.co.uk/news/world-middle-east-40276568; Ozgun Topak, ed., *New Authoritarian Practices in the Middle East and North Africa* (Edinburgh University Press: 2022).

[97] Export Control Joint Unit, "Interception and monitoring prohibitions in sanctions made under the Sanctions and Anti-Money Laundering Act 2018: technical guidance," 17 October 2023, https://www.gov.uk/government/publications/interception-and-monitoring-prohibitions-in-sanctions-technical-guidance/interception-and-monitoring-prohibitions-in-sanctions-made-under-the-sanctions-and-anti-money-laundering-act-2018-technical-guidance.

[98] James M. Acton, "Chapter 1: On the Regulation of Dual-Use Nuclear Technology," American Academy of Arts & Sciences, https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/4.

[99] Michael Laufer, "A. Q. Khan Nuclear Chronology," Carnegie Endowment for International Peace, 7 September 2005, https://carnegieendowment.org/research/2005/09/a-q-khan-nuclear-chronology.

was able to use this information to run a long-term nuclear proliferation ring, developing a network enabling the export of nuclear capabilities to countries including Iran.

The threat of nuclear proliferation has ensured that the field is heavily regulated and subject to several political and international agreements and treaties, setting it apart from other dual-use fields. In 2004 the UN Security Council introduced Resolution 1540, which was adopted to prevent non-state actors from accessing weapons of mass destruction (WMDs).[100] Following this, the Security Council introduced sanctions on Iran and North Korea to prohibit the export of nuclear materials. Sanctions were targeted at both governments as well as the external suppliers supporting nuclear programmes. The Treaty on the Non-Proliferation of Nuclear Weapons (NPT) was introduced in 1968 and has since acquired 191 member states.[101] Nuclear-weapon states are prohibited from assisting non-nuclear weapon states from acquiring or manufacturing nuclear weapons.[102]

The International Atomic Energy Agency (IAEA) plays an essential role in the implementation and inspection of safeguards. Non-nuclear-weapon states are required to complete a comprehensive safeguards agreement with the IAEA in order to comply with the NPT.[103] In 1995, the IAEA established the Incident & Trafficking Database, used to log and monitor incidents related to radioactive materials such as theft or sale.[104] 145 participating countries share incidents on a voluntary basis with the wider network, to strengthen national and global security through lessons learned and regular communication.

At a UK level, the Office for Nuclear Regulation (ONR) is the regulator for safety, security and safeguarding at all 35 nuclear sites, as well as for the transport of nuclear materials.[105] Safeguards employed by the ONR ensure that the UK is held accountable to the commitments made under international treaties. In an effort to provide better guidance on cybersecurity, the Civil Nuclear Cyber Strategy was rolled out in 2022.[106] The strategy was developed jointly with leaders from public and private sector civil nuclear organisations, the

[100] Acton, "Chapter 1: On the Regulation of Dual-Use Nuclear Technology."

[101] UN Office for Disarmament Affairs, "Treaty on the Non-Proliferation of Nuclear Weapons (NPT)," https://disarmament.unoda.org/wmd/nuclear/npt/.

[102] International Atomic Energy Agency, "The IAEA and the Non-Proliferation Treaty," https://www.iaea.org/topics/non-proliferation-treaty.

[103] International Atomic Energy Agency, "Safeguards agreements," https://www.iaea.org/topics/safeguards-agreements.

[104] International Atomic Energy Agency, "Incident and Trafficking Database (ITDB)," https://www.iaea.org/resources/databases/itdb.

[105] Office for Nuclear Regulation, "About ONR," https://www.onr.org.uk/about-us/.

[106] Department for Business, Energy & Industrial Strategy, *2022 Civil Nuclear Cyber Security Strategy*, May 2022, https://assets.publishing.service.gov.uk/media/627df8658fa8f53f9a15c1d5/civil-nuclear-cyber-security-strategy-2022.pdf.

ONR and the NCSC, and presents clear responsibilities for industry and regulators on how to better protect the nuclear cybersecurity landscape.

As well as regulation at an industry level, efforts are in place designed specifically to guide universities managing nuclear research. An example is the ATAS scheme,[107] which seeks to prevent individuals from gaining access to information which could later be used to build WMDs. Universities may also have links to radiation protection societies, which provide support and guidance for research institutions working with radioactive materials.[108] In relation to the security of hazardous materials, the Health and Safety Executive is accountable for providing guidance to institutions.[109] Aside from complying with regulation, strong physical and virtual access controls are in place and resilience is tested regularly, while retaining the facilitation of research as a top priority.

The nuclear sector has clear and established mechanisms in place to draw together cross-sector and international expertise. A shared risk perception and sense of collective responsibility encourages participation in agreements and networks. Given its history as a military technology, the international goal of non-proliferation has meant that nuclear safety and security is generally well-understood.

5.2 International case studies

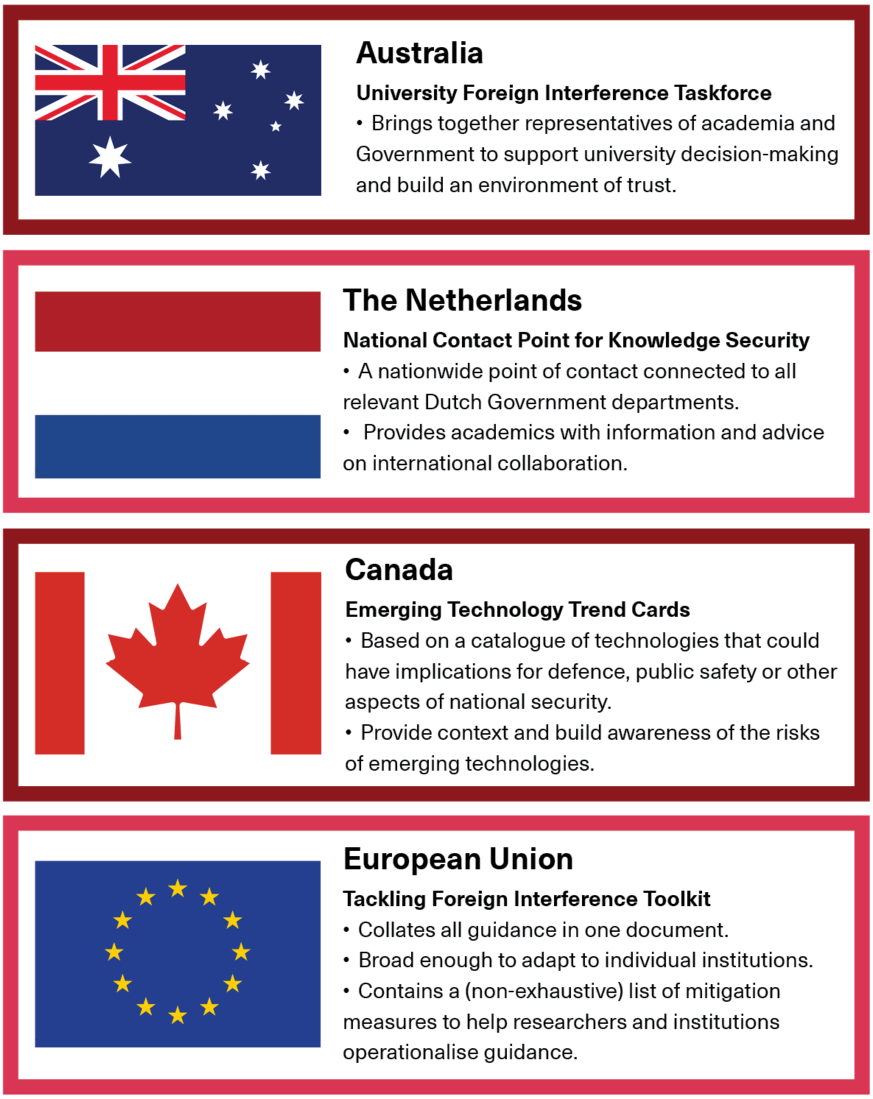
Figure 2 highlights cases of international best practice for strengthening academic resilience to state threats. From clear points of contact to easily accessible information and guidance, the common theme across these case studies is clarity of information for researchers and institutions. Academics are the ‘front line’ when it comes to the protection of research, and must be empowered to understand the threat landscape and mitigate risks.

[107] Foreign, Commonwealth & Development Office, “Academic Technology Approval Scheme (ATAS),” https://www.gov.uk/guidance/academic-technology-approval-scheme.

[108] Chung et al., *Nuclear Security within Academic and Research Organisations: A Handbook of Global Case Studies* (King’s College London Centre for Science & Security Studies: 2022), https://kclpure.kcl.ac.uk/ws/portalfiles/portal/170404255/NS_in_Academic_and_Research_Organisations.pdf.

[109] Health and Safety Executive, “Our mission and priorities,” https://www.hse.gov.uk/aboutus/our-mission-and-priorities.htm.

Figure 2. Examples of international best practice[110]



Australia

University Foreign Interference Taskforce

- Brings together representatives of academia and Government to support university decision-making and build an environment of trust.

The Netherlands

National Contact Point for Knowledge Security

- A nationwide point of contact connected to all relevant Dutch Government departments.
- Provides academics with information and advice on international collaboration.

Canada

Emerging Technology Trend Cards

- Based on a catalogue of technologies that could have implications for defence, public safety or other aspects of national security.
- Provide context and build awareness of the risks of emerging technologies.

European Union

Tackling Foreign Interference Toolkit

- Collates all guidance in one document.
- Broad enough to adapt to individual institutions.
- Contains a (non-exhaustive) list of mitigation measures to help researchers and institutions operationalise guidance.

Source: Authors' analysis.

[110] Government of the Netherlands, "Contact Point for Knowledge Security," https://english.loketkennisveiligheid.nl; Government of Canada, "Emerging Technology Trend Cards," https://science.gc.ca/site/science/en/safeguarding-your-research/guidelines-and-tools-implement-research-security/emerging-technology-trend-cards; European Commission, "Commission publishes a toolkit to help mitigate foreign interference in research and innovation," 18 January 2022, https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/commission-publishes-toolkit-help-mitigate-foreign-interference-research-and-innovation-2022-01-18_en; Australian Government, "Countering foreign interference," https://www.homeaffairs.gov.au/about-us/our-portfolios/national-security/countering-foreign-interference.

5.3 Maturity model

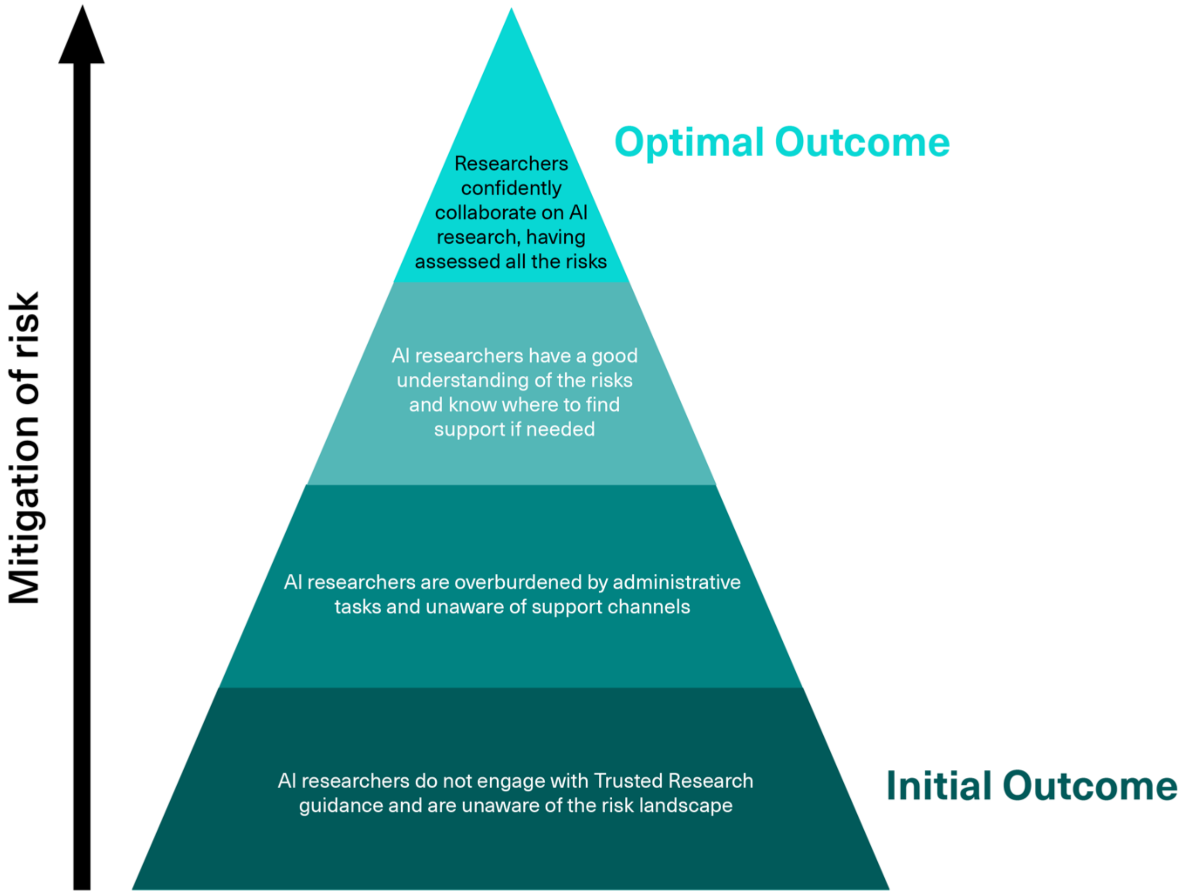
The below Maturity Model was developed during an internal workshop and is informed by primary and secondary data collected over the course of the research. The higher the level of maturity, the better the institution's resilience to research theft, acquisition and interference.

The model is intended to be used as a tool for academic institutions and AI researchers to prompt thinking and assess the maturity of their risk mitigation measures in response to the threat landscape. The drivers identified on the left-hand side are building blocks for reaching optimal maturity.

These drivers are not intended to be exhaustive, due to the dynamic nature of the risks to AI research security and the individual contexts applicable to each academic institution. The drivers broadly relate to the following themes:

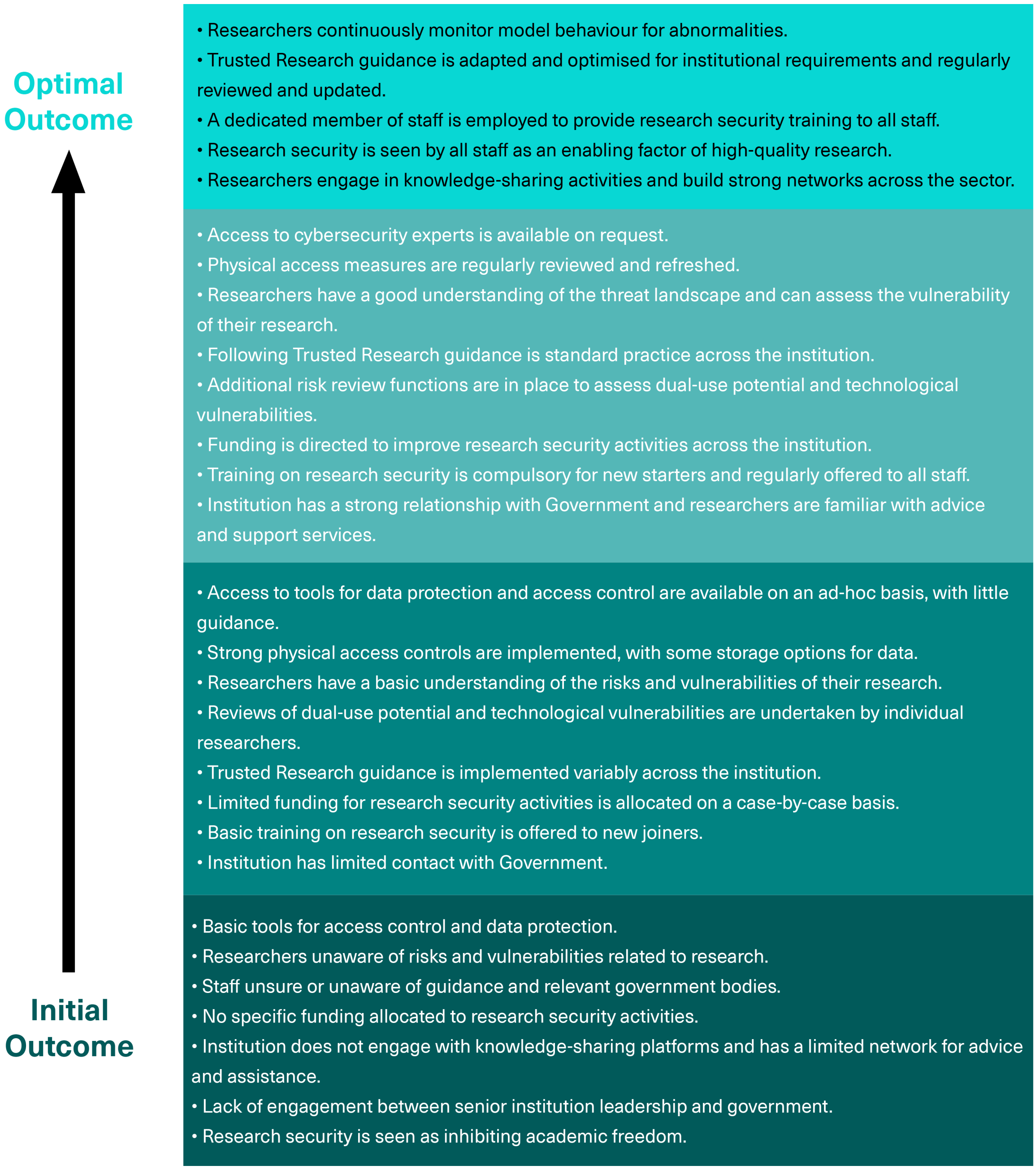
- Physical (including personnel) security and cybersecurity.
- Institutional policy.
- Funding and governance.
- Training.
- Culture.

Figure 3.1. Outline of maturity model


Mitigation of risk
Researchers confidently collaborate on AI research, having assessed all the risks
Optimal Outcome
AI researchers have a good understanding of the risks and know where to find support if needed
AI researchers are overburdened by administrative tasks and unaware of support channels
AI researchers do not engage with Trusted Research guidance and are unaware of the risk landscape
Initial Outcome

Source: Authors' analysis.

Figure 3.2. Inputs to maturity model



Optimal Outcome

- Researchers continuously monitor model behaviour for abnormalities.
- Trusted Research guidance is adapted and optimised for institutional requirements and regularly reviewed and updated.
- A dedicated member of staff is employed to provide research security training to all staff.
- Research security is seen by all staff as an enabling factor of high-quality research.
- Researchers engage in knowledge-sharing activities and build strong networks across the sector.

- Access to cybersecurity experts is available on request.
- Physical access measures are regularly reviewed and refreshed.
- Researchers have a good understanding of the threat landscape and can assess the vulnerability of their research.
- Following Trusted Research guidance is standard practice across the institution.
- Additional risk review functions are in place to assess dual-use potential and technological vulnerabilities.
- Funding is directed to improve research security activities across the institution.
- Training on research security is compulsory for new starters and regularly offered to all staff.
- Institution has a strong relationship with Government and researchers are familiar with advice and support services.

- Access to tools for data protection and access control are available on an ad-hoc basis, with little guidance.
- Strong physical access controls are implemented, with some storage options for data.
- Researchers have a basic understanding of the risks and vulnerabilities of their research.
- Reviews of dual-use potential and technological vulnerabilities are undertaken by individual researchers.
- Trusted Research guidance is implemented variably across the institution.
- Limited funding for research security activities is allocated on a case-by-case basis.
- Basic training on research security is offered to new joiners.
- Institution has limited contact with Government.

- Basic tools for access control and data protection.
- Researchers unaware of risks and vulnerabilities related to research.
- Staff unsure or unaware of guidance and relevant government bodies.
- No specific funding allocated to research security activities.
- Institution does not engage with knowledge-sharing platforms and has a limited network for advice and assistance.
- Lack of engagement between senior institution leadership and government.
- Research security is seen as inhibiting academic freedom.

Initial Outcome

Source: Authors' analysis.

6. Conclusion and Recommendations

AI research is a valuable asset and is vulnerable to state threats. There is a clear need to create more friction to deter state threat actors and build a more resilient AI research ecosystem, but tension exists between principles of academic freedom and research security.

The UK Government must continue to support academic researchers by providing accessible and clear guidance, building relationships with relevant bodies and working to address systemic problems such as the domestic AI talent pipeline. The academic community also has a responsibility to build and maintain a sector-wide culture of risk awareness and security-mindedness. Stronger incentives are needed to encourage consistent compliance with existing voluntary guidance and best practice.

6.1 Recommendations for the UK Government

1. The UK Government should commission a classified mapping of the AI research ecosystem in the higher education sector. This will provide the Government with a clearer overview of existing vulnerabilities and allow it to provide targeted support to institutions.

2. DSIT, with support from the NPSA, should provide regularly updated, direct guidance to research-intensive universities on international institutions deemed high-risk for funding agreements and collaborations. This will provide additional clarity to researchers.

3. The UK Government should provide DSIT with dedicated funding to grow the RCAT, a key conduit of information between the Government and academia. This would empower the RCAT to further invest in specialist technical staff and research capabilities that support academic due diligence.

4. DSIT should produce a white paper on the drivers of AI talent retention in academia. The AI Opportunities Unit should prioritise efforts to plug the AI skills gap and encourage young people to take up academic research roles. This would help the UK retain domestic talent and remain at the forefront of high-quality technological R&D.

5. The NPSA and the NCSC should engage more widely with UK-based publishing houses, academic journals and other research bodies to brief senior decision-makers on the threat landscape and offer support in developing policies tailored to research security.

6. The NPSA should declassify and publish case studies of relevant threats that have been intercepted or disrupted.

7. UKRI should provide specific grant funding opportunities for research security activities, aiming to encourage institutions to invest in research security training and the infrastructure to support this.

8. UKRI should standardise the T&Cs of its grants, providing researchers with greater clarity about the guidance and legal provisions they need to follow. These T&Cs should explicitly outline unacceptable behaviours and mandate audits of research security to incentivise best practice.

9. UKRI should set up a Research Security Committee to report directly to its board on the threat from state actors and risk mitigations, serving as an audit function for research security-related T&Cs.

6.2 Recommendations for academia

10. Academic institutions should deliver mandatory research security training (based on Trusted Research guidance) to new staff and postgraduate research students as a prerequisite of grant funding. This training should be accredited by the NPSA.

11. The academic sector should develop a centralised due diligence repository to document risks and inform decision-making on AI research partnerships and collaboration. This repository should be hosted by a trusted partner, such as UUK or UKRI.

12. Research-intensive universities should set up research security committees to help academics conduct risk assessments of their work on AI (and other critical technology).

13. Major AI journals and academic publishing houses should standardise pre-publication risk assessment for AI research, in line with existing processes for reviewing research ethics.

Annex A: Existing Legislation and Guidance

Legislation	Summary
UK Strategic Export Controls[111]	An export licence is required if software or technology is linked to items in the consolidated list of strategic military and dual-use items, including electronics, computers, and telecommunications and information security. AI is not listed as a specific item but would normally fall within one of the aforementioned subcategories.
National Security and Investment (NSI) Act 2021[112]	AI and dual-use technologies are both deemed sensitive to national security, with institutions being legally required to notify the Government about acquiring entities (e.g. a partnership or company) and assets in certain areas. Academic engagement with the Act relies on voluntary notification.
National Security Act 2023[113]	Under Section 39, the Home Secretary may impose restrictions on an individual’s work or studies if they are reasonably believed to be involved in “foreign power threat activity.” Section 2 on obtaining or disclosing trade secrets is also applicable to academia.

[111] Export Control Joint Unit, “Export controls applying to academic research,” 2 August 2024, https://www.gov.uk/guidance/export-controls-applying-to-academic-research; Export Control Joint Unit, “Consolidated list of strategic military and dual-use items that require export authorisation,” 3 April 2024, https://www.gov.uk/government/publications/uk-strategic-export-control-lists-the-consolidated-list-of-strategic-military-and-dual-use-items-that-require-export-authorisation.

[112] Cabinet Office, “Check if you need to tell the government about an acquisition that could harm the UK's national security,” 21 May 2024, https://www.gov.uk/guidance/national-security-and-investment-act-guidance-on-acquisitions.

[113] Home Office, “National Security Act 2023,” 20 January 2025, https://www.gov.uk/government/collections/the-national-security-bill.

Academic Technology Approval Scheme[114]	Applies to international students and researchers subject to UK immigration control and intending to study or research at postgraduate level in subjects where knowledge could be used in programmes to develop Advanced Conventional Military Technology, WMDs or their means of delivery.
Guidance Documents	**Summary**
NPSA Trusted Research Guidance for Academia[115]	Advice covers: • Conduct of due diligence. • Appropriate segregation of information, people and networks. • Cybersecurity measures to mitigate threats. • Checklist of good practices.
UKRI Trusted Research and Innovation Principles[116]	Advice includes: • Conduct of partner suitability assessments. • Robust information security management measures. • Implementation of collaboration agreements to manage sensitive data, intellectual assets and IP rights.
UUK Managing Risks in Internationalisation Guidance[117]	Advises institutions to: • Ensure sound governance. • Report annually on research security risk management. • Embed a security-minded culture.

[114] Foreign, Commonwealth & Development Office, "Academic Technology Approval Scheme (ATAS)."

[115] National Protective Security Authority, *Trusted Research Guidance for Academics*.

[116] UKRI, *UK Research and Innovation Trusted Research and Innovation Principles*, August 2021, https://www.ukri.org/wp-content/uploads/2021/08/UKRI-170821-TrustedResearchandInnovationPrinciples.pdf.

[117] Universities UK, "Security and risk: how universities can protect their research and people," 8 June 2023, https://www.universitiesuk.ac.uk/topics/funding-finance-and-operations/security-and-risk-how-universities-can.

Centre for Emerging Technology and Security

RESEARCH REPORT