SENATE

Title of paper: Proposal for NU Centre of Research Excellence in AI Safety

Main purpose of the paper: For decision

Presenter(s): Professor Matthew Grenby (PVC for Research and Innovation)

Date of paper: 17th June 2025

Purpose of the paper

This paper presents a proposed NU Centre of Research Excellence in AI Safety. It is intended that this NUCoRE will join a portfolio of 19 NUCoREs, which will enable the University to offer a coherent and distinctive narrative of our collective excellence in this distinct area of AI Safety research, education, and engagement with global reach. It is designed to enable the University to seize external funding in this growing area of research and innovation.

Relation to strategy and values

The consolidation of the University's multi-disciplinary research into NUCoREs is one of the 'transformative initiatives' in the University Research Strategy launched in October 2018. The NUCoRE will grow the University's research portfolio through upholding the aspirational values of excellence, creativity and impact.

Recommendations:

The documentation provides a snapshot of the current status of the NUCoRE, but approval is sought on grounds of future potential and trajectory.

Consultation to date (including any previous committee consideration and its outcome):

Community Workshop and external engagement - Autumn/ Winter 2024 followed by discussion with Research Strategy Implementation Group - Feb 2025

Faculty Executive Boards - endorsed – March – May 2025 (feedback enclosed)

University Executive Board – endorsed – May 2025 (feedback enclosed)

Appendix A - NUCoRE Proposal: Centre for AI Safety

NUCoRE name	Centre for AI Safety (CAIS)
Lead(s)	Prof Rajiv Ranjan
Lead Faculty	SAgE

1. What? *Indicative 150 words*

Establishing the Centre for AI Safety (CAIS), in collaboration with the National Edge AI Hub[1,2], is a significant initiative to place Newcastle University as a leader in AI Safety Research. AI Safety has become a significant concern in the UK and at the international level.

The establishment of the Centre for AI Safety at Newcastle represents a bold step toward addressing one of the most critical challenges of our time. It will not only enhance the university's research excellence and reputation, but also contribute to ensuring that the benefits of AI are realised safely, ethically, and equitably, both nationally and globally. This initiative aligns with Newcastle University's mission to deliver cutting-edge research with meaningful impact and its commitment to leading on issues of vital importance to society.

CAIS is grounded in cross-university partnership and aims to significantly enhance Newcastle University's research excellence, impact and reputation, foster a dynamic research culture, stimulate new learning initiatives and partnerships, and leverage existing Artificial Intelligence systems and applications strengths. By nurturing early career researchers and promoting cross-faculty research, CAIS will support the University in developing the next generation of research leaders.

In addition, by leveraging the resources, expertise, and partnerships developed through the National Edge AI Hub, CAIS will act as a catalyst to unlock the University's latent potential. The Centre will mobilise underutilised capabilities within Science, Agriculture and Engineering (SAgE), Faculty of Medical Sciences (FMS), and Humanities and Social Sciences (HaSS), creating synergies that enhance Newcastle University's societal and economic impact. It will also position the University as a key player in shaping the future of AI governance, safety standards, and public trust in emerging technologies.

Externally, as part of the Centre for AI Safety's strategic expansion, Lenovo has agreed in principle to establish a Joint National AI Safety Lab at Newcastle University in collaboration with the Centre for AI Safety and the National Edge AI Hub. This Lab will enable high-impact, contracted research projects, with Lenovo providing state-of-the-art AI hardware and the Centre offering world-leading expertise in AI safety, training, and assurance. We expect the total investment by Lenovo to be £200k minimum. In addition, Singapore Design University has also agreed in principle to establish the Newcastle-SDU Joint Centre in Trusted and Safe AI, which will facilitate contracted research projects across Southeast Asia and strengthen Newcastle's position as a global leader in safe and trustworthy AI.

In a climate of financial upheaval, CAIS's plan to undertake contracted research will generate new revenue streams for the University. Moreover, by attracting top talent, increasing research funding and grants, collaborating with national and international partners, and influencing public policy and regulatory frameworks, CAIS will elevate Newcastle University's standing and impact on society and policy both within the UK and internationally. Ultimately, through these strategic activities and aims, we are united in our mission to contribute substantially to Newcastle University's vision of advancing knowledge, providing creative solutions, and solving global problems.

Distinctness of the CAIS

The proposed Centre for AI Safety at Newcastle University will be highly complementary to the existing Newcastle University Centres of Research Excellence (NUCoREs) by addressing the cross-cutting challenge of AI safety across a wide range of disciplines, including work, education,

[1]According to our current correspondence with UKRI, these AI hubs are scheduled for a 10-year lifespan (much like the UKRI Quantum Hubs which were recently extended).

[2] https://edgeaihub.co.uk/

healthcare, culture, digital governance, and energy. These examples are illustrative of the Centre's broad relevance. CAIS will provide a dedicated research ecosystem to ensure that these innovations are safe, reliable, and aligned with ethical and regulatory frameworks.

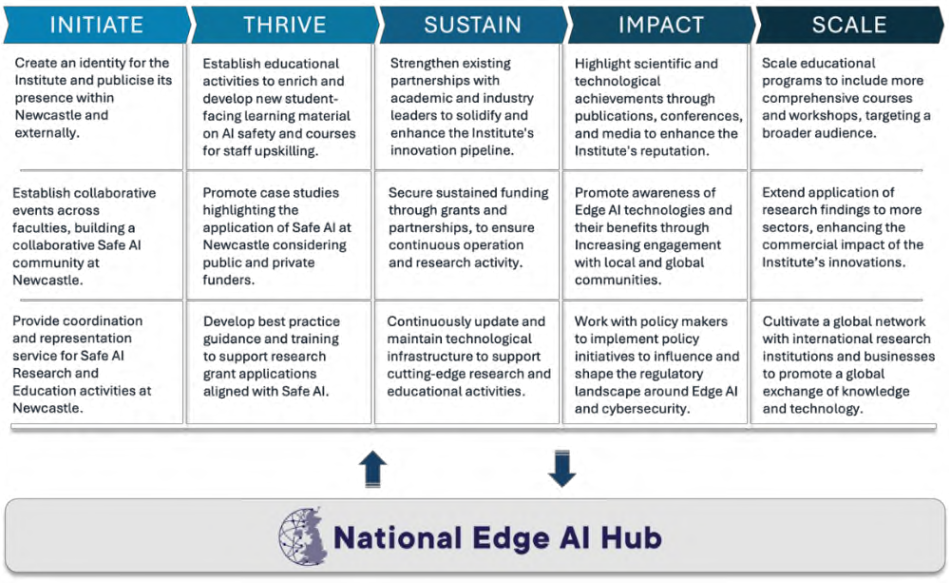
We assert that Artificial Intelligence Safety is a distinct and critical discipline and has no overlap with existing NUCoREs in Data/AI and Cyber. Data science is focused on extracting insights from information. Foundational AI is centred on developing core models and algorithms. Cybersecurity aims to protect systems and networks. In contrast, AI Safety is about designing and building autonomous decision-making systems that can operate safely, ethically, and reliably in real-world environments, particularly in high-stakes domains like medicine, where safety risks are life-critical.

Unlike the National Edge AI Hub, which focuses specifically on the cyber-resilience of AI algorithms in distributed edge computing environments, CAIS will adopt a broader and more foundational scope. It will address safety risks across all levels of AI systems, from the development of large foundational models to their deployment in real-world industrial and societal contexts. In doing so, CAIS will actively collaborate with NUCoREs and the National Edge AI Hub to embed safety-by-design principles into AI-driven solutions, reinforcing Newcastle's leadership in AI Safety research and policy development.

Execution plan for the next 5 years

These initiatives position Newcastle University as a leader in advancing safe and effective AI integration across academia, industry, and society. The overall execution plan of the Centre is shown in Figure 1. The five-year execution plan reflects our commitment to delivering impact through strategic collaboration, research excellence, and practical implementation. We recognise the importance of maintaining flexibility within this plan to ensure we stay ahead of rapid developments in AI and continue to shape best practices in this evolving field.

Figure 1. CAIS 5-year Plan

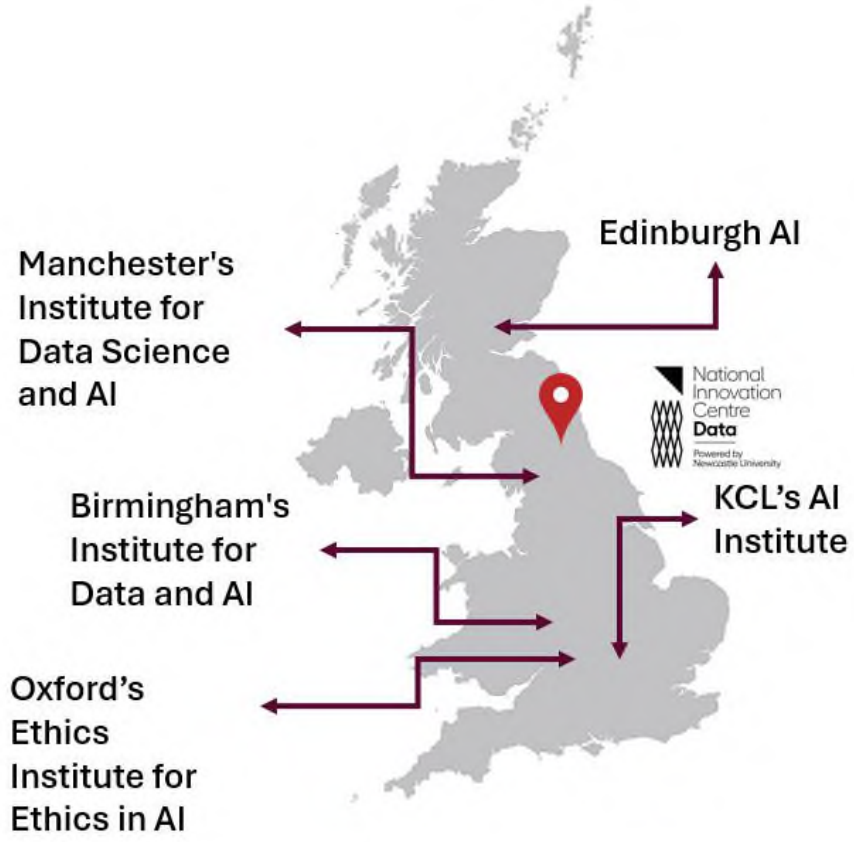


2. Why? *Indicative 150 words*

The establishment of CAIS is not just a strategic initiative but an urgent response to the rapid advancements in AI technology and the economic and social risks they pose. These advancements' far-reaching implications across various sectors and society underscore the immediate and pressing need for AI safety.

Rapid advancements in AI technology and their implications across various sectors and society have prompted UK's leading universities to invest in AI institutes. These institutes promote cross-disciplinary collaboration and interactions among university researchers to advance the field of AI. For example, KCL launched the AI Institute[3] in 2022, and Oxford launched the Institute for Ethics in AI[4] in 2021. Similarly, other leading universities (see Figure 2), such as Cardiff, Birmingham, and Leeds, have data science institutes offering similar cross-disciplinary research, innovation and education platforms.

Figure 2. National Data Science Initiatives

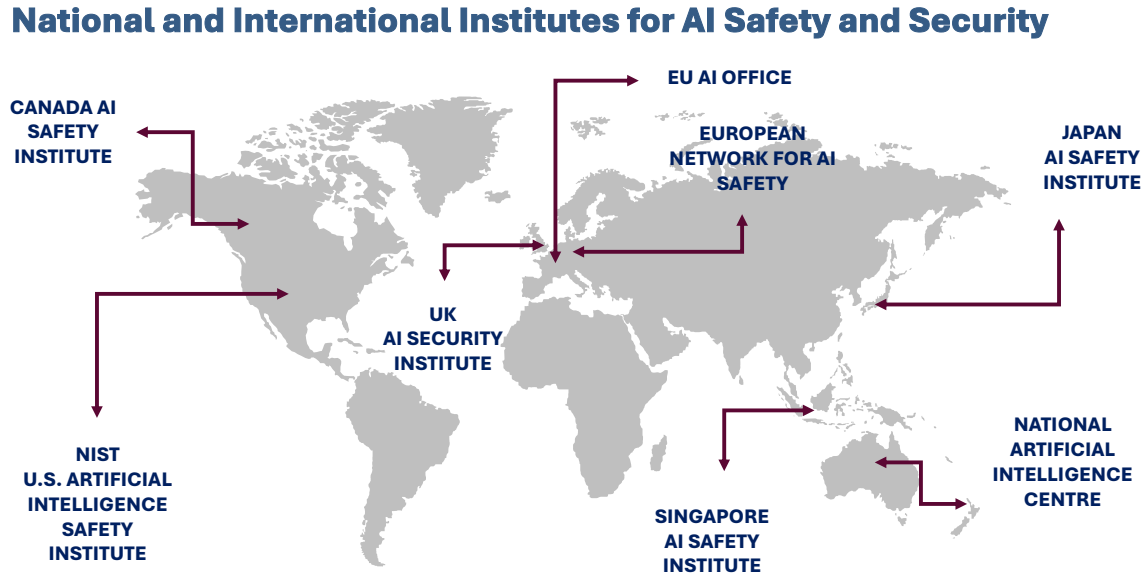


However, one of the most pressing needs today is to ensure AI advancements are safe for our society. Therefore, to harness the advantages of advances in AI while minimising the potential for misuse and harm, the UK Government has launched the AI Security Institute (AISI)[5] with follow-up investments in 9 AI research hubs, including our National Edge AI Hub, hosted and directed by Newcastle University.

[3] https://www.kcl.ac.uk/ai
[4] https://www.oxford-aiethics.ox.ac.uk/
[5] https://www.aisi.gov.uk/

Furthermore, AI safety is a global priority (see Figure 3), as demonstrated by investments in AI safety across the world (US[6], Japan[7], Europe[8], and Australia (National AI Centre)[9]).

Figure 3. International AI Safety and Security Institutes



While Newcastle and other UK universities have invested in data science initiatives (such as NICD at Newcastle) and other AI institutes (such as Ethics for AI at Oxford), *we advocate the first university-based AI Safety Centre with a holistic focus on safety*. For example, the primary focus of AISI is to develop and conduct evaluations on advanced AI systems (such as large language models) and be advisory to the Government in creating effective AI governance and regulation based on these evaluations, while the Responsible AI Hub is focused on the societal and ethical impact of AI.

On the other hand, the focus of CAIS will be on preventing harm by ensuring the safe operation of AI systems (e.g. digital twins for crisis resilience and mitigation, autonomous vehicle safety, fraud detection systems, smart agriculture, smart manufacturing, and energy security). Non-functional attributes of such AI systems, including reliability, availability, assurance, privacy, integrity, equality, and transparency, are not considered by any of the above initiatives in a holistic way.

Establishing CAIS will thus uniquely position Newcastle University as a pioneer in the field, coordinating and embedding AI safety according to nationally agreed-upon standards in research, development, AI deployment and benefits realisation. This initiative aligns with the UK AISI to address the needs of the Newcastle University research and education community, and the North East's public and private sectors, and beyond. By partnering with the UK's AI Security Institute[10], a Department for Science, Innovation, and Technology directorate, CAIS will achieve significant strategic benefits.

[6] https://www.nist.gov/aisi
[7] https://aisi.go.jp/
[8] https://www.enais.co/
[9] https://www.csiro.au/en/work-with-us/industries/technology/national-ai-centre
[10] The partnership will be fostered through the National Edge AI Hub and UKRI [personal communication with Dr Kedar Pandya, Executive Director, Cross-Council Programmes].

3. Consultation and Development plan *Indicative 150 words*

The development of the Newcastle Centre for AI Safety will build on the momentum generated by the National Edge AI Hub, leveraging its established synergies with Newcastle University's faculties, SAgE, FMS and HaSS, as well as key industry partners. *Different from the National Edge AI Hub*, the Centre will focus on addressing safety concerns in general AI contexts, ensuring its impact spans various sectors, including healthcare, smart cities, and sustainable development.

The demand for establishing the Newcastle Centre for AI Safety stems from multiple key events hosted by the National Edge AI Hub. During its national launch in May 2024, attended by Newcastle University academics and industry leaders, stakeholders highlighted the need to expand beyond edge-specific AI cyber-resilience to include broader AI safety concerns across all faculties. This sentiment was echoed during the online launch of the National Edge AI Hub later that month. The momentum continued to build during the Newcastle Centre for AI Safety community event in November 2024, where a diverse group of participants reinforced the importance of a dedicated Centre for AI safety (community feedback available on request).

We also held an ESRC IAA Innovation Forum in January 2025 to debate the real-world challenges and risks in business and law associated with AI deployment, and a London event in February 2025 to discuss how to transform edge AI innovation into tangible solutions, as well as a number of webinars for research and practitioners. Such events aim to engage the community, ensuring the Centre remains relevant and responsive to evolving needs and opportunities. This ongoing engagement will continue in order to guide the Centre's development, positioning it as a national leader in interdisciplinary AI safety research and practice.

Key Development Milestones:

- Hosted interdisciplinary community event (November 2024) with participation from SAgE, FMS, HaSS, and industry partners.
- Developed collaborations with industry leaders, via National Edge AI Hub events, to address real-world AI safety challenges.
- Incorporated ongoing community feedback from National Edge AI Hub and ESRC IAA Innovation Forum events into the Centre' roadmap.
- Expanded the scope of AI safety research to encompass general AI applications across diverse sectors.

4. Who? *Indicative 150 words*

To ensure the Centre for AI Safety receives dedicated attention, a robust governance structure will be implemented including a senior leadership team and dedicated theme leads for each work package. This distributed leadership and governance model will guarantee representation, continuity, focus, and alignment with the Centre' vision.

Prof. Rajiv Ranjan, as the Director of the National Edge AI Hub and a globally recognised leader in AI safety and resilience, is uniquely positioned to lead the Centre. His extensive expertise in managing large-scale, multidisciplinary projects ensures the Centre' goals align with national and international AI safety priorities. Prof. Ranjan will provide strategic leadership critical to securing high-impact collaborations, driving research excellence, and positioning Newcastle University at the forefront of AI safety innovation.

The National Edge AI Hub will provide tangible administrative and business development support (in-kind) through its Hub Operations Manager, Impact Manager and Project Coordinator roles. We expect that through synergies but also dedicated activities 25% of these posts will contribute towards the Centre's operations. In due course, we also plan to extend the project coordinator role which at the moment is a Part-Time role into a full-time one and provide more dedicated support for the Centre. In addition, the Hub will offer research support through projects undertaken by its PDRAs/RSEs.

When it comes to stakeholder engagement it will provide access to its Independent and Industrial Advisory Boards. In addition to the above, the Hub also plans to spend £150k to setup National AI Safety Laboratory that will be directly linked to the Centre' objectives. The Hub is already in discussions with external partners such as Lenovo who are interested in investing in such a facility. We expect the investment to be £200k minimum.

Although difficult to quantify precisely, the National Edge AI Hub is expected to contribute approximately £200k per year in in-kind support—amounting to £1 million over the planned five-year period. The establishment of the Centre for AI Safety will enable Newcastle researchers to seamlessly access £2.5 million in flexible funding currently held by the Hub, along with a national collaboration network of 12 universities and 60 industry partners. In total, up to £2.5 million in flexible funds can be leveraged by Newcastle researchers to develop nationally recognised REF 2029 impact case studies in partnership with the Hub's academic and industrial stakeholders.

Governance & Interdisciplinarity

We will appoint new Application Area Theme Specialists (AATSs) from across the faculties, with representation from SAgE, HaSS, and FMS. These individuals will play a key role in aligning AI safety-related research and education efforts within their respective faculties, working closely with the Centre's Research and Education Directors and the corresponding Directors from Schools across FMS, SAgE, and HaSS. In parallel, CAIS will designate leads for education and impact to strengthen cross-faculty collaboration. The National Edge AI Hub theme leads will work in partnership with AATSs to support the coordinated delivery of research and education activities. Workload management will align with Newcastle University's EDI Charter and be implemented in consultation with the Faculty PVC and Heads of Unit who have already agreed to provide support.

Moreover, in communication with UKRI, it has become evident that there will be more funding towards AI and Health going forward. Therefore, as the CAIS evolves, we expect more theme specialists, especially from FMS and HASS, to join the Centre who can help spearhead submissions to such grant calls.

5. How? *Indicative 150 words*

The Newcastle Centre for AI Safety will deliver impact through seven dynamic work packages, tackling public engagement, capacity building (grant acquisition), and impact monitoring, alongside cutting-edge research in algorithmic robustness, real-time AI safety, ethical and regulatory frameworks, and human-AI collaboration. This interdisciplinary framework will inform a bold, comprehensive approach to advancing AI safety. The proposed work-packages (WPs) and sub-themes will encourage interdisciplinary research, engagement, and impact across the three faculties: SAgE, FMS, and HaSS.

Work-Package 1: Public engagement, education, and workforce development (Lead(s): Research Impact Director, Responsible Innovation and Education Director, and National Skills Lead**)**

- Objective: Foster societal understanding of AI safety and prepare the next generation of leaders in safe AI development.
- Sub-themes:
    1. AI Safety programs (SAgE, FMS, HaSS):
        - Developing interdisciplinary PGT/CPD courses and training for students and professionals in AI safety.
    2. Public awareness campaigns on AI risks and benefits (HaSS):
        - Collaborating with policymakers, educators, and community organisations to enhance AI literacy.
    3. Upskilling healthcare, STEM, and Social Science professionals (FMS, SAgE, HaSS):
        - Equipping professionals across sectors with tools to navigate the challenges and opportunities of AI safely.

Work-Package 2: Interdisciplinary grant applications and funding (Lead: Centre Director**)**

- Objective: Facilitate collaborative research proposals to secure funding from UKRI, the European Commission, and industry partners, aligning with AI safety goals.
- Sub-themes:
    1. Developing competitive grant proposals (SAgE, FMS, HaSS):
        - Target UKRI's Strategic Priorities Fund, particularly in areas such as Digital Security by Design, Responsible AI, and Future of Health.
        - Leverage European Commission's Horizon Europe funding calls, such as "Human-Centred and Ethical AI" and "Trustworthy AI for Future Societies."
    2. Strengthening industry partnerships (SAgE, FMS, HaSS):
        - Engage with industry players to co-develop grant applications for Innovate UK funding programs like Smart Grants and Commercialising Connected and Autonomous Mobility.
        - Develop partnerships to address challenges identified in Catapult Network Initiatives, such as AI applications in manufacturing and health tech.
    3. Exploring European and international funding opportunities (SAgE, FMS, HaSS):
        - Apply to international schemes such as the EIC Pathfinder for high-risk, high-reward research in AI safety.

- Build global consortia to address calls from the OECD AI Policy Observatory and other global AI safety-focused programs.

Work-Package 3: Impact and policy influence evaluation / REF 2029 Impact (Lead (s): Responsible Innovation and Education Director and REF2029 Impact Director**)**

- Objective: Develop frameworks to measure AI safety outcomes and contribute to policymaking at national and international levels.
- Sub-themes:
    1. AI Safety metrics and benchmarks (SAgE):
        - Establishing technical benchmarks for measuring AI safety and performance.
    2. Impact evaluation of AI in public health and safety (FMS):
        - Developing methods to assess the real-world impact of AI in healthcare and social care.
    3. Influencing national and international AI policy (HaSS):
        - Engaging with government bodies and international organisations to influence AI safety standards and regulations.
    4. Developing inter-disciplinary REF 2029 impact study (all):
        - Lead a cross-faculty initiative to develop a robust REF2029 impact study, showcasing interdisciplinary advancements in AI safety.

Work-Package 4: Scientific excellence in AI Safety for critical systems (Lead: Theme Lead1**)**

- Objective: Develop robust, scalable, and safe AI systems for critical applications across healthcare, infrastructure, and public safety.
- Sub-themes:
    1. Safe AI for Edge and Cloud Computing (SAgE):
        - Fault-tolerance, redundancy, and resilience mechanisms for cloud-based AI applications. *Please note that edge computing-based mechanisms, which are fundamentally different from cloud-based mechanisms, will be provided by the National Edge AI Hub.*
    2. AI for patient safety in digital health (FMS):
        - Ensuring the safety and reliability of AI-driven diagnostics, decision-making tools, and autonomous healthcare devices.
    3. Risk perception and societal trust in AI (HaSS):
        - Understanding societal perceptions of AI risks and developing frameworks to enhance public trust in AI adoption, acceptance and diffusion.

Work-Package 5: Ethical AI and socio-technical governance (Lead: Theme Lead2**)**

- Objective: Address fairness, accountability, transparency, and ethical decision-making in AI systems, ensuring alignment with societal values.
- Sub-themes:
    1. Fairness and bias mitigation in AI algorithms (SAgE, FMS, HaSS):
        - Techniques to eliminate algorithmic bias in autonomous systems and machine learning models.
    2. AI ethics in clinical applications (SAgE, FMS, HaSS):

        - Developing ethical frameworks for AI tools in personalised medicine, clinical trials, and telemedicine.
    3. Policy and regulatory frameworks for Safe AI (SAgE, FMS, HaSS):
        - Researching governance structures, human rights implications, and regulatory mechanisms for responsible AI.

Work-Package 6: AI in decision-making for sustainable development (Lead: Theme Lead3**)**

- Objective: Leverage AI for sustainable and safe societal development, addressing challenges like climate change, energy use, and resource management.
- Sub-themes:
    1. AI for smart and sustainable urban Systems (SAgE):
        - Safe AI applications in urban planning, energy optimisation, and environmental monitoring.
    2. AI for public health and pandemic response (FMS):
        - Deploying safe AI to monitor public health trends and respond to global health crises.
    3. Ethical AI for social equity and justice (HaSS):
        - Ensuring AI applications promote equity, reduce inequalities, and prioritise marginalised communities.

Work-Package 7: Human-centric AI and interaction design (Lead: Theme Lead4**)**

- Objective: Focus on designing AI systems that prioritise human safety, well-being, and ease of use.
- Sub-themes:
    1. Human-AI interaction and cognitive safety (SAgE):
        - Designing user-centred AI interfaces that minimise errors and maximise safety in human-AI collaboration.
    2. AI for mental health and well-being (FMS):
        - Safe and ethical AI applications for mental health monitoring, support, and therapy.
    3. Impact of AI Adoption (HaSS):
        - Examining the cultural and psychological effects of AI systems on diverse user communities.

Appendix B – NUCoRE proposers feedback to Faculty Executive Board feedback

Feedback by HASS FEB

1. **Such a project would benefit from physical space on the campus and as this is a cross-Faculty project any proposal would need to go to Estates Portfolio Board**

Mitigations: Indeed, this is an important priority. We are already exploring options for establishing the National AI Safety Laboratory lab with investment from the EdgeAI Hub and external partners such as Lenovo. Creating a dedicated facility of this kind could have significant implications for both the institute and the university. Beyond generating income, it could also open opportunities for offering complementary solutions, such as AI safety certifications. Securing appropriate space will be a key enabler for this vision, and we would be more than happy to liaise with the Estates Portfolio Board to identify a long-term solution that aligns with the scale of our ambitions.

2. **To navigate this transition effectively, NU will have to establish robust guidance to manage the integration of AI tools. Additionally, retraining the entire workforce will be essential to ensure that these tools are used efficiently, safely, and ethically. The Institute can therefore also serve a crucial internal purpose, assisting staff across the university in the adoption of the tools as we transition to being an AI-enabled organisation.**

Mitigations: In addition to sharing the University's expertise in AI safety and contributing positively beyond our institution, we are also committed to fostering meaningful advancements within the University itself. We seek to support colleagues in exploring and adopting AI tools thoughtfully and responsibly, ensuring they enhance our work while aligning with ethical principles. We can explore dedicated training and collaborative initiatives, in order to create an environment where AI is not only understood but actively leveraged to improve our own practice.

3. **Namely the 5-year plan, which neatly packages intended outcomes into years. I am assuming this means these activities starting up and continuing across the subsequent years (though that's not how it's represented here). I think the plan as it stands however could be conflated, and that we need to be working much earlier on the later outputs. AI isn't going to wait for us to catch up, and it is moving so quickly, we may have to be much more on the front foot to keep up.**

Mitigations: While a forward-looking plan provides a structured outline, it is not intended to be rigid or strictly sequential. Instead, key activities will need to start earlier and progress in parallel to ensure we remain ahead of developments rather than merely reacting to them. To that end, we will refine the plan to better reflect this agility, ensuring that later-stage outputs are initiated earlier and that we are continuously working towards our long-term goals. The Institute is a long-term commitment, and we are dedicated to adapting our approach as necessary to maintain momentum and leadership in this rapidly evolving space.

4. **The proposed AI Safety Institute claims it will look at 'multiple disciplines, including healthcare, cybersecurity, digital governance, and energy'. However, this raises the question of why education is not at the top of this list, in particular HE which is poised to undergo significant transformation in the coming years due to the widespread integration of AI across all workstreams.**

Mitigations: We have taken this into consideration and have added work and education as indicative/potential areas of focus. These areas are not exhaustive, and we remain open to expanding our scope as needed to address the evolving landscape of AI's impact across various sectors. Our commitment is to ensure that the AI Safety Institute remains adaptable and responsive to emerging trends and challenges.

Feedback by SAgE FEB

1. **The recommendation of SAgE FEB was that it was happy to approve the Centre but as the Centre for AI Safety and not as an Institute and that it should genuinely be "no cost".**

Mitigations: Expecting a grant ambition to be realised at virtually no cost is unrealistic. Instead, the key consideration should be the return on investment. The proposal seeks in-kind, no-cost support to ensure that those involved have the necessary resources to establish the Institute and deliver its mission effectively. Unlike other NUCoREs, the EdgeAI Hub is making a substantial investment in this joint endeavour, not only through staff time but also through direct financial contributions. Additionally, the Hub is actively working to secure external funding, such as the investment from Lenovo for the National AI Safety Laboratory. This demonstrates a strong commitment to ensuring the Institute's success while leveraging strategic partnerships to enhance its sustainability.

2. **Designation as an Institute rather than a NUCoRE/Centre. This issue prompted considerable discussion. There was a concern that a lot of work had been done over a number of years to build the NUCoRE "brand" which was aimed at bringing coherence to a space that had become complex with inconsistent branding across centres, institutes etc.**

Mitigations: **We have mutually agreed to call this initiative as the "Centre for AI Safety".**

3. **There was also concern about the potential for confusion with the government's national AI Safety Institute (albeit the name has been changed recently to the AI Security Institute). This confusion was not considered to be a purely semantic issue, but could have practical implications for the drive to attract significant external investment on the back of the Government's AI strategy. Retaining the "Centre" as the Centre for AI Safety would help distinguish the NUCoRE from other entities nationally and internationally.**

Mitigations: As the point above highlights, the AI Safety Institute has now been renamed the AI Security Institute, which reduces the risk of confusion and creates an opportunity to position ourselves more distinctively in this space. This distinction allows us to establish a clear identity while reinforcing our unique contributions to AI safety. In turn, we agree that strategic positioning is crucial for attracting external investment and ensuring the long-term sustainability of the Institute by generating new resources and forging key partnerships.

4. **Is the Centre genuinely "no-cost"? It was noted that support for the Centre amounted to a total of 3.15 FTE across the university (2.25 FTE from SAgE).**

Mitigations: This has been addressed at the beginning of this sub-section. Workload management will align with Newcastle University's EDI Charter and be implemented in consultation with the Faculty PVC and Heads of Unit. The University is not expected to contribute any cash in real terms.

5. **While much of the staff FTE is attributed to the EPSRC AI Safety Hub some elements are not (e.g. Finance and Project Management Support indicated in Figure 4 of the proposal). No FTE is identified for these positions in Table 1. It was not clear where this support would come from especially in current circumstances with reduction in size of both PS and academic staff.**

Mitigations: Financial and project management responsibilities will be handled by the Hub Manager and Project Coordinator. We recognise the need for dedicated support and plan to extend the Project Coordinator's contract to a full-time position in due course, ensuring greater capacity within the Institute itself.

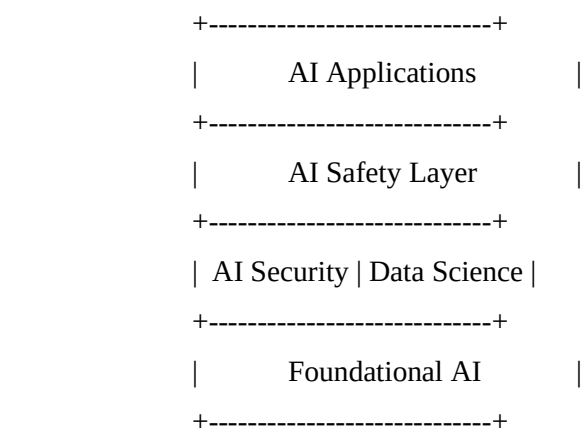
Feedback by FMS FEB

1. **Felt to overlap too closely with existing NUCOREs in Data and Cyber.**
    - **Case for unique area vs tangential extension of discipline not made.**
    - **Lack of justification for a standalone entity.**
    - **Feeling it should be a strand with an existing NUCORE.**

Mitigations 1.1: AI Safety is Not Data Science or Cybersecurity

We fully acknowledge the committee's concern about perceived overlap with existing NUCOREs in *Data* and *Cyber*. However, we respectfully but firmly assert that Artificial Intelligence Safety is a distinct discipline:

- Data science is about extracting insights from information.
- Foundational AI is about core models and algorithms.
- Cybersecurity is about protecting systems and networks.
- Artificial Intelligence Safety is about building autonomous decision-making systems that operate safely, ethically, and reliably in real-world settings, including medicine, where the risks are life-critical.



AI Safety is not a subfield of either discipline (Data science, Foundational AI and/or Cybersecurity); it requires its own scientific frameworks, ethical debates, regulatory strategy, and translational infrastructure. The misconception itself emphasises why there is a need for a standalone focus on AI Safety. Around the world and in the UK, there has been AI safety institutes have been established to understand and develop AI Safety frameworks, tools and techniques. Several funding opportunities/calls are solely dedicated to fund AI Safety and AI Safeguarding AI topics.

Mitigations 1.2: Structure and Governance

While we welcome collaboration with NUCOREs in Data and Cyber, AI Safety requires:

- A dedicated research identity
- Its own advisory board (including FMS)
- The ability to host funded doctoral cohorts, research fellows, and innovation programmes tied to AI governance and risk

We propose close links, not subsumption. The Centre can act as a node of cross-NUCORE integration, while also standing as an academic brand and strategic asset in its own right.

Mitigations 1.3: We Are in the Middle of an AI Revolution

We are at a technological inflection point. Just as universities responded to the genomics boom (Centre for Life) or the internet age with strategic investment in new structures, we must now respond to the AI revolution:

- The emergence of foundation models, generative AI, and autonomous agents is reshaping healthcare, education, policy, and society. These models are powerful but opaque, and their deployment in health settings brings significant risks around fairness, robustness, explainability, and control.
- The current system of embedding AI research within data science or cyber units is no longer adequate for the scale, speed, and societal stakes of modern AI.
- Universities that treat AI as "just another tool" risk being left behind. Those who lead in safe, ethical and human-centered AI will shape the next decade of policy, practice, and impact.

A standalone Centre for AI Safety positions Newcastle as a National Hub for Safe, Deployable, Interdisciplinary AI—not only advancing research, but actively shaping policy, regulation, and innovation.

Mitigations 1.4: The Cost of Inaction: What If We Don't Build This?

Failing to establish an AI Safety Institute carries significant risks:

- Strategic marginalisation: Without a dedicated AI Safety identity, Newcastle will lose relevance in AI Safety policy, funding, and collaboration, especially as government and industry increasingly seek focused AI institutions.
- Fragmented research: By forcing AI Safety to live within other NUCOREs, we dilute its focus, slow interdisciplinary engagement, and undermine its ability to attract world-class talent and partners.
- Missed economic opportunity: The Northeast has a real opportunity to be a testbed for safe AI in health and public services. Without this institute, we will forfeit our chance to lead, and watch other universities and regions assume that mantle.

2. **Concerns around the nomenclature of Institute:**
    - **This is founded on a single project grant/collaborative grouping and does not have equity to the large line-management Institutes within FMS.**
    - **Seen as a retrograde step and would set a precedent with other**

Mitigations: In response to your comments regarding the use of the term "Institute", we fully acknowledge the importance of maintaining consistency with the established naming conventions for large, line-managed Institutes within FMS. We understand the concerns that the proposed title may imply a similar governance or structural model, which is not the intention of this initiative.

Accordingly, we have revised the name of the proposed entity to Centre for AI Safety to more accurately reflect its nature as a collaborative, interdisciplinary grouping anchored by a major project grant and external partnerships, rather than a formal line-management structure. We believe this change aligns with University conventions and avoids setting any unintended precedents.

Thank you once again for your valuable input. We look forward to continued collaboration with FMS colleagues as the Centre develops.

3. **Lack of true cross-faculty representation**

Mitigations 3.1: Cross-Faculty and Regional Integration (FMS and Beyond)

We recognise the concern about limited FMS representation. We will take action to ensure deep, structural integration with FMS, including:

FMS Strategic Alignment:

FMS School	**AI Safety Applications**
Medicine	Clinical AI validation, diagnostic model safety, predictive analytics
Psychology	Human-AI trust, decision-making under AI support
Pharmacy	AI in drug safety, automation in prescription systems
Population Health	Bias audits, equity models, public health surveillance

Named representatives from each of these Schools will be invited to serve on the Institute's Advisory Group, ensuring co-creation of strategic priorities.

Mitigations: **3.2*: Regional Collaboration***

- We will align closely with Daiser, Health Call, AHSN NENC, and regional NHS Trusts, who see safe, validated, trustworthy AI as essential to future service transformation.
- These partners need a translational centre that can assess, monitor, and validate AI before real-world use—a function that cannot be embedded generically within data, AI or cyber NUCOREs.

Mitigations 3.3: We are actively working towards organising the first ideas factory workshop that will be dedicated to health services and not just AI related to medical treatment in September. Hence, we expect to be able to build more bridges with FMS and colleagues interested in working with us. Needless to say, that we will welcome any specific suggestions as to people with whom we can liaise in the first instance.

Feedback by University Executive Board

Noted:

- Received recommendations for the establishment of a NUCoRE in AI Safety.
- The proposed NUCoRE in AI Safety was intended to enable the University to offer a coherent and distinctive narrative of collective excellence in this distinct area of AI Safety research, education, and engagement with global reach.
- Noted that there had been extensive and detailed consultation and discussion of the proposed NUCoRE in AI Safety at Faculty Executive Boards and at URIC.
- Reflected on the potential of the NUCoRE in AI Safety to contribute to teaching. A new Master's programme was under discussion.
- It would be good to hear about how researcher and leadership development could be part of the Centre's work, ie. mentoring early career colleagues with the Centre.
- A 'checkpoint' should be built in at around 2 or 2.5 years after the Centre's establishment, to ensure that things are progressing as planned.

- **Agreed:**
- Executive Board endorsed the recommendation that Senate approve the establishment of a NUCoRE in AI Safety.