SEPTEMBER 2024

JAKOB MÖKANDER
HELEN MARGETTS
ROBERT TRAGER
KEEGAN MCBRIDE
NITARSHAN RAJKUMAR
MARIE TEO

Getting the UK's Legislative Strategy for AI Right

TONY BLAIR
INSTITUTE FOR
GLOBAL CHANGE

Contents

A joint position paper

01

Executive Summary

Artificial-intelligence systems are powerful technologies that bring significant economic and social benefits. However, their development and use are coupled with ethical, social and security risks. Reaping the benefits of AI will require incentives for innovation, robust engineering and safety practices, and effective, proportionate regulation. Properly designed, such regulation can accelerate AI adoption by unlocking investments in digital infrastructure and increasing public trust in the technology.

The UK has so far adopted a pro-innovation[1] and sector-specific approach[2] to AI regulation, relying on existing regulators to oversee the use of AI within their domains. This approach has many merits, as it is adaptable and grounded in context. But it does leave gaps. Many regulators lack resources, powers and technical AI expertise, leaving concerns regarding algorithmic bias, data privacy and misinformation insufficiently addressed. Further, as the capabilities of advanced general-purpose AI systems (or "frontier AI"[3]) continue to advance, a more centralised and proactive approach may be needed to identify and manage emerging public-safety risks.[4]

Against this backdrop, the government is drafting an AI bill to address public-safety concerns associated with frontier AI while advancing innovation to support its growth agenda. Recent media reports[5] indicate that the Department for Science, Innovation & Technology (DSIT) intends the bill to do two things: 1) make companies' existing voluntary AI safety commitments legally binding, and 2) turn the AI Safety Institute (AISI) into an arm's-length government body.

Getting the AI bill right is vital to strengthen the UK's position as a global leader in AI innovation, use and safety. Yet this will be a complex task. There is no consensus around which risks are most pressing to address. There is also major uncertainty about what capabilities future AI systems will have, how to best evaluate such systems and address their risks, and thus what legally binding requirements are needed. Questions about enforcement are further complicated by geopolitical factors and the information asymmetry between AI developers and governments.

At the same time, improving the safety of frontier-AI systems is only one piece of the puzzle. Equipped with a strong mandate, the government has the opportunity to develop an overarching legislative strategy for AI that supports its five missions.[6] It should seize this opportunity by ensuring that sector-specific regulators have adequate resources, powers and coordinating functions to perform their duties and by providing a clear roadmap for addressing regulatory gaps. However, while a comprehensive legislative strategy for AI is needed, not everything has to go into this first bill. Moreover, any AI regulation should work hand in hand with public policy to support investments in AI research and development, in particular provision of public resources for compute.[7]

Taking the government's recent announcements as a starting point, the Tony Blair Institute for Global Change has convened key stakeholders from industry, academia and civil society to produce this joint position paper, laying out considerations and recommendations to inform the planned AI bill and the equally important task of getting the UK's overarching legislative strategy for AI right.

RECOMMENDATIONS

1. **The UK government should develop a comprehensive legislative strategy for AI, maintaining and strengthening its sector-specific approach to AI regulation.** This includes addressing regulatory gaps and resolving regulatory tensions. Existing regulators require significantly increased funding, AI expertise and powers to perform their duties. Some shared regulatory capacity may be needed to improve transparency and accountability across the AI value chain. A clear roadmap for implementing the overarching legislative strategy during this parliament should be delivered alongside any bill that focuses narrowly on AI safety.
2. **If an initial narrow bill is pursued, its aim should be to build regulatory capacity and advance scientific understanding of the cross-cutting public-safety risks posed by frontier AI.** New tools, methods and institutions are needed to effectively evaluate, monitor and report on frontier-AI systems' dual-use capabilities, limitations and impacts and manage their risks. A narrow bill focusing on AI safety is only one initial part of the broader action required on AI governance and should be explicitly framed as such.
3. **AISI should be made an independent, technical and non-regulatory arm's-length body.** Its tasks should include: 1) advancing the science of

evaluating and monitoring frontier-AI systems, 2) promoting the international standardisation of AI safety tools and methods, and 3) supporting third-party assessments of the safety of frontier-AI systems and the adequacy of AI developers' safety practices. AISI should not become a regulator as that would threaten its status as a trusted, independent technical body.

4. **A narrow bill focusing on frontier-AI safety should be flexible, allowing for incremental implementation and the delegation of technical decisions.** Given the rapid evolution of AI capabilities and safety practices, technical details – for example regarding thresholds, requirements, model access, evaluation metrics and reporting formats – should largely not be codified in primary legislation but instead delegated to competent and sufficiently resourced authorities. Building on AI developers' existing voluntary commitments, the government should take a principles-based[8] approach that allows for flexibility and innovation in safety practices.[9]
5. **In drafting both a narrow bill and in its overarching legislative strategy for AI, the UK should build on its global leadership to promote greater international alignment on AI regulation.** To the extent that is possible and accounting for geopolitical factors, the UK should strive to ensure consistency with regulatory developments and standard-setting processes in the United States and the European Union. It should also collaborate on safety research and advance global technical standards and best practices through the international network of AI safety institutes.
6. **Any binding regulations should be accompanied by incentives for relevant actors to comply;** this may involve clarifying existing regulators' responsibilities as part of the overarching legislative strategy. A new regulator focusing on safety issues that cut across sectors but are specific to frontier AI may eventually be needed. As AI safety standards and scientific understanding of risks are still nascent, there is debate about whether this should be done in the near term or delayed to a separate bill, potentially with interim powers being granted to the relevant secretaries of state. The government should provide a clear timeline for any subsequent bill to maintain public trust and regulatory certainty for developers.

The remainder of this position paper expands on our recommendations and discusses the merits and limitations of the various policy options.

02

Approach and Framing

Recommendation 1: *The government should present a comprehensive legislative strategy for how to strengthen the existing sector-specific regulatory approach for AI. This should include increased funding to existing regulators in the upcoming Budget, as well as a clear roadmap for addressing known regulatory gaps with respect to broad AI risks through targeted legislation during this parliament. Only alongside these actions can a bill focused on improving the safety of frontier-AI systems be justified as addressing a regulatory gap.*

The UK's approach to AI regulation has many advantages. It maintains a pro-innovation[10] stance while delegating responsibility to sector-specific regulators. This is sound, given that a multitude of complementary policy responses is required to manage the full range of risks AI systems pose, from product liability to public safety.[11] It is also responsive to the sociotechnical nature of many AI risks that surface in different use cases. Sector-specific regulators with relevant domain expertise are best placed to understand and address the impact of AI on their sectors.

While the UK's sector-specific approach merits building upon, it needs strengthening. Many existing regulators lack the resources and expertise to address AI-related risks such as data-privacy violations, intellectual-property infringements and algorithmic discrimination. Each regulator will need to address those risks in their own way. Sponsoring departments should be responsible for improving their capacity and proposing new legislation as needed. Adding to individual departments' efforts, the government should substantially increase AI-related funding for regulators in the upcoming Budget.

The government should also outline an overarching legislative strategy for addressing known regulatory gaps through new legislation and present a clear timeline for implementation during this parliament. Some areas where AI is used – for example, edtech – have no regulator. Other AI risks, such as misinformation, impact many sectors, from health care to online safety. Having "AI officers"[12] in each department could improve coordination across sectors. Further, addressing use-case-specific risks requires transparency and accountability across the AI value chain, which may call for obligations on actors who are not clearly in the scope of existing regulation.

One critical gap relates to the public-safety risks[13] that highly capable general-purpose AI systems may pose. According to the International Scientific Report on the Safety of Advanced AI,[14] dual-use capabilities[15] can amplify cyber- and biosecurity risks. While these risks are currently low, the government has a responsibility to anticipate and respond to future societal risks. As of today, however, it lacks the tools to identify, assess and manage public-safety risks from frontier AI.[16]

A proactive bill focusing on frontier-AI safety could help address that gap by advancing scientific understanding of AI risks, incentivising relevant actors across the AI value chain to strengthen their safety processes and improving the government's regulatory preparedness. Such a bill would complement – not compete with – the work of existing regulators.

A bill focusing on frontier-AI safety must not preclude the need for rigorous thinking in other areas of AI governance.[17] The government's overarching legislative strategy for AI should include steps towards: 1) building regulatory capacity in specific sectors, and 2) providing robust assurance for other parts of the AI value chain. Further, while frontier-AI systems may pose security risks, we caution against framing a narrow bill or AISI's role solely in national-security terms.

Finally, terminology matters. Should the government proceed with a narrow scope, the name "AI bill" may be misleading, since it suggests more comprehensive action across the spectrum of AI risks and the AI value chain. The "frontier-AI safety bill", the "AISI bill" or similar would be more appropriate labels. It should be noted that "frontier AI" is itself a poorly defined term. This position paper only refers to it for continuity with the language used in previous communications from the UK government and AISI as well as by AI developers during the AI Seoul Summit 2024.

03

Scope and Aim

Recommendation 2: *A narrow bill – as proposed by the government – should focus on contributing to public safety in a manner complementary to the UK's sector-specific approach. Its aim should be to build regulatory capacity and advance the science of evaluating and managing frontier-AI safety risks. This means improving transparency and accountability around frontier-AI models' capabilities and limitations and imposing binding risk-management and reporting obligations on well-resourced AI developers in line with their existing voluntary commitments.*

A narrow bill focusing on frontier-AI safety should strengthen the government's regulatory capacity and address regulatory gaps. This can be done in several ways, for example by advancing the science of evaluating AI systems, improving transparency and accountability throughout the AI value chain, and promoting international standardisation of AI safety practices.

A key justification for the focus on frontier-AI safety is to minimise future uncertainty.[18] Current frontier-AI systems pose limited cyber-, bio- and national-security risks.[19] Hence, a restrictive regulatory approach is not justified today on those grounds. However, well-resourced actors are investing billions of dollars in the development of ever-more capable AI systems, and there is uncertainty about the capabilities that these will have. A proactive approach is needed to manage future AI risks. Making companies' voluntary frontier-AI safety commitments legally binding and putting AISI on a statutory footing are steps in the right direction.

To address public-safety concerns, even a narrow bill must consider its implications for the AI value chain.[20] AISI has so far developed tools for the evaluation of models before their release. But AI safety is not an isolated model property.[21] Many risks depend on *how* models are integrated into a broader system, *who* is using the system and for *what* purposes. In addition to model evaluations and assessments of AI developers' safety practices,[22] post-deployment monitoring and incident reporting are needed to identify risks that arise in real-world settings and should fall within the bill's scope.[23] While AISI need not conduct such monitoring, it can help develop tools and standards for it.

Initially, the bill should limit its scope to focus on well-resourced developers of highly capable AI systems as they are best positioned to take on additional compliance burdens. Low-resource AI development should initially be exempt. A more extensive regime would be infeasible for such a diffuse technology, given the limited resources the government can deploy, and is currently undesirable because of the restrictions it would place on innovation by smaller actors.

There is still debate concerning the safety of open versus closed models.[24] To date, the impact of open release has been positive, with significant benefits for research and innovation.[25] The bill should take care not to undermine that impact. However, open release may become a problem as AI capabilities increase and the cost of training models falls. Restrictions on openness will require a high standard of evidence for justification – which only reinforces the importance of gaining improved visibility of current and future AI capabilities.

The bill should define its key terminology carefully.[26] The government must provide clear legal definitions of terms like “frontier AI” or “general-purpose AI”, “public safety” and “open source”. In doing so, it should build on existing efforts. For example, Article 3 of the EU AI Act provides a provisional definition for general-purpose AI.[27] Similarly, a legal definition for open-source AI could build upon that of the multistakeholder Open Source Initiative.[28]

A final point on scope. When drafting the bill, the government should not only think about how to keep people safe *from* AI but also *with* AI. It is important to develop defensive AI capabilities, to protect against cyber-threats and support resilience in line with the Cyber Security Strategy.[29] This should be delivered through existing bodies like the National Cyber Security Centre[30] and be kept consistent with sector-specific regulation such as the Online Safety Act.[31] AISI should also continue to catalyse work in this area through its Systemic AI Safety programme.[32]

04

The Role of the AI Safety Institute

***Recommendation 3:** AISI should be put on a statutory footing as an independent technical body advising the government, existing sector-specific regulators and the public. Importantly, AISI should not become a regulator. As an arm's-length body, its main roles should be to: 1) advance the science of evaluating and monitoring frontier-AI systems, 2) promote international standardisation of AI safety tools and methods, and 3) support third-party assessments of the safety of frontier-AI systems and the adequacy of AI developers' safety practices.*

AISI is respected due to its technical expertise, like the non-regulatory National Institute of Standards and Technology in the US or the National Physical Laboratory in the UK. Accordingly, it should be put on a statutory footing as an independent arm's-length technical body,[33] with operational freedom and limited political control. This is important to ensure a continued collaborative and productive relationship with industry. It should not become a regulator, as this trusted position would be undermined.[34]

Focusing on technical work,[35] AISI's function should be to advance the science of how to evaluate and monitor AI systems; quantify and report on AI capabilities, limitations and public-safety risks; design engineering safety practices for the development of advanced AI systems;[36] and enable independent third-party assessments of developers' safety practices. As the first and most well-funded body in a nascent global network of AISIs,[37] the UK's AISI should take a leading role in shaping international AI safety standards and promoting regulatory harmonisation.[38]

AISI should not conduct all the evaluations that fall within a narrow AI bill's scope itself. The responsibility to ensure that AI systems are legal and safe falls on the companies that develop or deploy them and might best be served through an AI assurance market (which the government could enable through Advanced Market Commitments).[39] AISI should be free to conduct model evaluations and safety-framework assessments of developers insofar as this furthers its mission to advance scientific understanding. But its main role should be to develop the standards needed to make meaningful assessment, monitoring and enforcement possible.

A sociotechnical approach will be needed to address many AI risks.[40] Ideally, information about model limitations should inform the design of downstream applications, and contextual knowledge about downstream harms should inform the design of model evaluations. To enable such information interfaces, AISI will need to work with other organisations such as UK Research and Innovation[41] to support the development of robust evaluation tools and methods across the value chain,[42] including at the governance, model and application layers.[43] AISI's Systemic AI Safety Fast Grants,[44] which focuses on these issues, should be expanded and given increased funding.

Evaluations conducted at different parts of the AI value chain are most effective when connected to structured procedures. The system access, tooling and infrastructure that AISI builds – and the information it gathers from developers – could thus be useful to other regulators as a sort of common regulatory capacity.[45] AISI could also contribute to such capacity in other ways, for example by facilitating collaboration on key AI issues, conducting risk mapping and horizon scanning, and sharing human and technical resources.

To enable these synergies, AISI should be given a mandate to share relevant information with sector-specific regulators and the public (for clearly defined purposes, with appropriate levels of transparency that respect trade secrets). This would not only allow developers of downstream applications and society at large to identify risks from specific models but also inform sector-specific regulators' efforts to hold AI developers who are not abiding by their safety commitments to account.

When putting AISI on a statutory footing, the government should heed lessons from the recent past. The Centre for Data Ethics and Innovation – now called the Responsible Technology Adoption Unit – was supposed to become an independent body but never did. Its effectiveness has been reduced as a result. The government would be well-advised to avoid repeating this mistake with AISI. The National Data Guardian could be a good model to draw on;[46] it has statutory independence from the Department of Health & Social Care and focuses purely on producing guidance but has no regulatory powers, as those sit with the Information Commissioner's Office.[47]

Finally, as it is put on a statutory footing, AISI should become more publicly transparent about the amount of funding it receives and how its resources are allocated. That would allow for clearer prioritisation between funding its work and that of other entities, such as regulators. The government has so far allocated £10 million[48] to jumpstart existing regulators' capabilities – compared with £400 million[49] to AISI until the end of the decade. There is also a strong case for exploring industry contributions to AI safety testing and monitoring procedures, as is standard practice in many other regulated sectors.

05

Delegation and Implementation

Recommendation 4: *The government should not rush the bill but instead prioritise getting the details right. Technical questions and policy design choices should be delegated to relevant departments and regulators or, where appropriate, independent expert bodies like AISI. The bill should outline a flexible, incremental implementation plan that allows for feedback and adaptation over time. While UK legislation typically works through delegation, resisting the temptation to mandate too much centrally is especially important in the case of AI development.*

AI research is rapidly advancing, and many technical AI-governance problems remain open, including key issues such as how to conduct effective evaluations.[50] A bill focused on frontier-AI safety will require an incremental implementation plan flexible enough to adapt to evolving AI capabilities and safety practices. This includes delegating aspects of key technical decisions (for example on thresholds,[51] model access[52] or evaluation standards[53]) to relevant departments, regulators and expert bodies. Specifically, DSIT should engage in thorough consultations with AISI, existing regulators and the wider ecosystem to get the details of the bill right.

One open question concerns what legally binding requirements (if any) should be put on AI developers. The voluntary frontier-AI safety commitments made by 16 major companies during the AI Seoul Summit 2024 constitute a good starting point.[54] These included pledges to: 1) assess the risks posed by frontier-AI systems across their lifecycle, 2) set out thresholds for intolerable risk, 3) articulate how risk mitigations will be implemented to keep risks within those thresholds, and 4) establish processes to follow if models exceed them.

These voluntary commitments are sensible and, importantly, supported by industry. However, not all of them can easily be translated into legally binding requirements. Moreover, many leading AI developers have signed up to several different sets of voluntary safety commitments – including those of the White House[55] – and are already subject to other binding regulations including the US Executive Order on AI.[56] The UK government should not limit itself by codifying any existing voluntary commitments into primary legislation. Rather, a thorough and open-ended process to develop effective and proportionate requirements is called for.

Drawing lessons from other regulated industries, there are many options available to foster transparency and accountability. These include measures like protected whistleblowing services, improved risk governance,[57] mandatory audits,[58] bug bounties for safety vulnerabilities[59] or requiring companies to publish regular statements on the efficacy of safety practices. When developing legally binding requirements, the UK should also incorporate lessons from ongoing international regulatory efforts, like the development of the EU's General-Purpose AI Code of Practice.[60]

There is also debate about how to identify AI models worthy of scrutiny.[61] Compute thresholds (for example, 10^25 floating-point operations in the EU AI Act[62] or 10^26 in the US Executive Order) have been used as an imperfect proxy for risk. While compute thresholds offer benefits like measurability and correlation with a broad swathe of AI capabilities,[63] they also have limitations.[64] Another possible approach is the $100 million training-cost threshold used in California's AI bill, SB 1047,[65] which targets the most powerful AI models without burdening startups and academics. Yet monetary thresholds also have limitations as the cost of training models may reduce over time and small models focused on acquiring harmful capabilities are possible.

Dynamic composite thresholds that combine multiple imperfect proxies for AI capabilities – like compute and training cost – and are updated over time will likely be required.[66] How these will be defined and updated remains an open question. Therefore, it would be premature to define thresholds in primary legislation in the near term, so questions regarding such technical details should be subject to extensive expert consultations and delegated to responsible departments, regulators and relevant technical bodies to ensure adaptability.

Taken together, the government must not rush the drafting of AI legislation as that could generate excessive or poorly targeted regulation. This would undermine the UK's status as an AI innovator and the collaborative and productive relationship between AISI and AI developers. Promisingly, there are ways of balancing pace and rigour. For example, the Automated Vehicles Bill successfully fuses clarity of intention and flexibility in implementation.[67] A frontier-AI safety bill should be crafted in a similar fashion.

06 International Collaboration

***Recommendation 5**: The government should seek to align both the bill and its overarching legislative strategy with international AI regulation. Regulatory fragmentation is not only bad for AI safety – as it can lead to a race to the bottom – but also for business and trade. One important goal should be to avoid duplication, such as where AISI repeats assessments previously conducted elsewhere. The UK should especially seek regulatory compatibility and consider mutual recognition with the US and the EU.*

In drafting and implementing the AI bill and its overarching legislative strategy, the government should build on the global leadership it has already shown in AI safety and pave the way for increased international collaboration and interoperability.[68] That might mean having technical bodies such as AISIs in different countries sharing best practices, notifying each other of risks, and standardising risk thresholds and evaluation techniques.

Where an AI model or developer has undergone public-safety testing in another jurisdiction that is also a signatory to the Bletchley Declaration[69] and Seoul Declaration,[70] the UK need not repeat the process. This will allow more effective use of government resources and lower the regulatory burden on industry. It will also address developers' concerns about espionage or the leaking of sensitive information if they grant model access to multiple countries. Sector-specific regulators should continue assessing the product-safety risks of AI in their respective domains. In health care, for instance, some national regulators such as the US Food and Drug Administration (FDA) and UK Medicines and Healthcare products Regulatory Agency (MHRA) have already entered "mutual recognition procedures" for medical devices.[71]

One important question concerns what kind of model access[72] is necessary for different types of evaluations.[73] This question has both technical[74] and geopolitical[75] dimensions. Some evaluations may require more than API-based structured access (for example, to IP such as training data and model weights).[76] However, clear justifications are needed for greater levels of access, and challenges exist around how to maintain security and the confidentiality of sensitive IP.

The UK needs to be realistic about its negotiating power and ability to shape firms' behaviour. Most AI development occurs outside the UK. Therefore, measures that only slow down or restrict model deployment in the UK do little to improve AI safety. Seeking too much access from firms could also create an unwelcome precedent where other countries seek similar levels of access in ways that undermine commercial IP security.

To the extent that is possible, the UK should seek to influence and align with relevant AI regulations, especially the EU AI Act Codes of Practice and the US Executive Order. Finding common ground will be difficult but not impossible, as demonstrated by the recent agreement around the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.[77] AI safety standards are key in harmonising international AI regulations and enabling mutual recognition of public-safety evaluations.[78] There are some existing AI standards (for example, the standards developed by the ETSI's Securing AI group[79] and certain International Organization for Standardization/International Electrotechnical Commission standards[80]) that could form the basis of a UK regulatory testing framework. Emerging recommendations around model-weight cyber-security should also be accounted for in the bill.[81] However, these are just starting points, and the UK AISI will have to develop and promote frontier-AI safety standards through the global network of AISIs.

International considerations also impact the timing of a UK bill. The outcome of the US election in November could affect the bill's reception by US companies and policymakers. Concurrently, UK legislation could positively influence the development of the EU AI Act Codes of Practice, which will be completed by April 2025. An ideal time for the introduction of the bill might be early 2025 around the AI Action Summit in France[82] – after the US election but before the April deadline for the EU's Codes of Practice.

To summarise, the UK government should develop legislation that is consistent with UK values and – where possible – borrows from and aligns with emerging AI safety regulations and best practices in other jurisdictions. The UK should seek to play a leading role in helping guide and shape the same. Increased international collaboration on AI with the EU and the US will be particularly important, including at a technical level with shared safety practices, risk taxonomies, evaluation infrastructures and reporting mechanisms.[83]

07 Enforcement

Recommendation 6: *Any binding regulations should be accompanied by incentives for relevant actors across the AI value chain to comply. AISI should monitor and publicly report compliance but not enforce it. Enforcement should primarily be ensured by clarifying the scope of existing regulators. However, this is a task for the overarching legislative strategy for AI, not a narrow bill. Eventually, a new regulator with a focus on frontier-AI safety may be needed. In the meantime, the relevant secretaries of state could be given new powers to apply penalties for non-compliance based on evidence provided by AISI, while implementing safeguards against politicisation.*

Voluntary AI safety commitments should be made legally binding only alongside clear incentives for relevant actors – including AI developers – to comply. But how to ensure compliance remains an open question. Experience from other sectors has shown that voluntary codes of practice are often ignored in the absence of monitoring and enforcement. To ensure public trust and safety, the government will need levers to intervene if there is sufficient risk of harm, even if it expects to use these only in rare circumstances.[84] A rigorous debate about the nature of these levers is required.

The government should be pragmatic about the feasibility of enforcing rules on a new and developing technology that society is still working to understand. It should also be mindful of its flourishing AI-innovation ecosystem and realistic about geopolitical considerations like the UK's ability to unilaterally enforce frontier-AI regulations. The absence of similar developments in the US is another reason for not regulating AI models at this stage but focusing on their use.

In theory, potential enforcement mechanisms include but are not limited to: 1) imposing fines or criminal penalties for non-compliance, 2) delaying the wide release of AI models, or 3) restricting access to UK resources (for example, UK Biobank) or government contracts.

In practice, however, fines have limited impact on companies' behaviour, and it would be hard for the UK to globally block model releases unless such actions are internationally coordinated. Further, each enforcement action would need to be considered in the context of the trade-offs in terms of its impact on the

UK as a location for AI research and innovation. Still, the bill should include enforcement powers to signal to frontier-AI developers the importance of compliance and build justification for the enforcement of stronger regulation in the future.

AISI itself should not become a regulator, and it should not enforce compliance with emerging safety practices. In the first instance, enforcement should be ensured by clarifying – or, if needed, expanding – the responsibilities of existing regulators. A separate regulatory function focusing on frontier-AI safety may eventually be needed. This could be established as a new entity or through adapting an existing regulator. But it is not clear that this is required imminently, given the uncertainty over the future public-safety risks AISI is set to focus on.

The government's overarching legislative strategy for AI should address the long-term question of enforcement. Setting up a new regulatory function is likely to take several years. If this is the government's intention, delays to its realisation should be minimised to provide regulatory stability. Until responsibilities are clearly delegated to existing regulators or a new regulator established in a subsequent bill, new powers for applying penalties for non-compliance with legally binding requirements could be accorded to the relevant secretaries of state. While this option has some precedent, we caution against politicising the enforcement of AI safety regulations.

In parallel, the UK should work with the EU and the US to harmonise frontier-AI safety regulations and enforcement levers. A good first step would be to focus on procedural obligations, like demanding AI developers follow emerging best practices for model testing and evaluation from AISI and other international bodies, and transparency obligations, for example disclosing AI model's capabilities and limitations to governments.

08

Conclusion

Having won a large parliamentary majority in the recent general election, Labour has a strong mandate to rebuild Britain. Properly designed, used and regulated, AI can help the government deliver on its five missions,[85] from accelerating economic growth to building an NHS fit for the future. The AI Opportunities Action Plan is a step in the right direction, as is the government's recognition that fostering AI innovation and ensuring good AI regulation go hand in hand.[86]

Our support of the the UK's current approach is grounded in a careful assessment of the current context. First, AISI has a significant talent pool and is a linchpin of the UK's global leadership in AI safety. Its independence should be increased. Second, the voluntary commitments made by AI developers may ultimately prove insufficient to ensure public safety.[87] The desire to stipulate binding regulations is thus understandable.[88] Third, the government has so far resisted calls to expand the scope of an initial bill. A narrow focus on frontier-AI safety can be justified, not because other concerns are less important, but because they can be dealt with by other means, including granting increased funding and powers to existing sector-specific regulators.

That said, getting the details right is crucial. The government must not rush any bill as that may lead to excessive or poorly targeted regulation. That would, in turn, risk producing a framework that quickly becomes out of date, undermines AISI's strengths, or reduces the UK's attractiveness as location for AI research and innovation. Further, widely established standards for safe frontier-AI development and deployment have yet to emerge. Getting the bill right will require careful iteration and international alignment. The government should opt for an incremental approach, flexible enough to adapt to advances in AI research and safety practices and allow lessons from other jurisdictions' regulatory efforts to be incorporated.

Finally, a narrow bill focusing on the safety of frontier-AI models only makes sense as part of an overarching legislative strategy to strengthen the UK's sector-specific approach to AI regulation. If the government proceeds with a narrow bill, it should label it as such (for example, as a "frontier-AI bill" or an "AISI bill"). The government should also articulate clearly what regulatory gaps the bill seeks to address and how it supports existing regulators in performing their duties. Any narrow bill should also be accompanied by: 1) greatly increased AI-related funding for existing sector-specific regulators, and 2) a timeline for addressing other AI risks through targeted legislation during this parliamentary session.

ABOUT THE AUTHORS

- Jakob Mökander is director of science & technology policy at the Tony Blair Institute for Global Change and an international fellow at the Digital Ethics Center, Yale University.
- Helen Margetts is professor of society and the internet at the Oxford Internet Institute, University of Oxford, and director of public policy at The Alan Turing Institute.
- Robert Trager is co-director of the Oxford Martin AI Governance Initiative, and senior research fellow at the Blavatnik School of Government, University of Oxford.
- Keegan McBride is a departmental research lecturer in AI, government and policy at the Oxford Internet Institute, University of Oxford.
- Nitarshan Rajkumar is a PhD candidate in AI at the University of Cambridge, and was previously senior advisor to the secretary of state for DSIT.
- Marie Teo is senior advisor for Global Government Engagement at the Tony Blair Institute for Global Change.

Editor's note: Due to the large number of contributors, co-authorship does not imply agreement with every point made in the position paper.

The authors also wish to thank Guy Ward-Jackson and Tom Westgarth for their research in support of this position paper.

CONTRIBUTORS

This joint position paper has been developed through broad stakeholder consultation across industry, academia and civil society. Contributors include:

- Alexander Babuta, Director, Centre for Emerging Technology and Security
- Ben Robinson, AI Policy Manager, The Centre for Long-Term Resilience
- Gina Neff, Deputy CEO, Responsible Ai UK
- Jack Clark, Co-founder, Anthropic
- Markus Anderljung, Director of Policy and Research, Centre for the

Governance of AI

- Max Fenkell, Head of Government Relations, Scale AI
- Mihir Kshirsagar, Policy Clinic Lead, Princeton Center for Information Technology Policy
- Owen Larter, Director of Public Policy, Microsoft
- Rebecca Stimson, Director of Public Policy, Meta

Editor's note: Contribution does not equal endorsement of all the points made in the position paper.

Endnotes

[1] https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper - :~:text=It%20is%20an%20approach%20that,scale%2Dup%20and%20compete%20internationally

[2] https://www.gov.uk/government/publications/regulators-strategic-approaches-to-ai/regulators-strategic-approaches-to-ai

[3] The UK government has previously defined "frontier AI" as "highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models". Today, that mainly includes large language models and other generative-AI systems.

[4] https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international%5Fscientific%5Freport%5Fon%5Fthe%5Fsafety%5Fof%5Fadvanced%5Fai%5Finterim%5Freport.pdf

[5] https://www.ft.com/content/ce53d233-073e-4b95-8579-e80d960377a4

[6] https://labour.org.uk/change/mission-driven-government/

[7] https://institute.global/insights/tech-and-digitalisation/state-of-compute-access-how-to-bridge-the-new-digital-divide

[8] https://arxiv.org/abs/2407.07300

[9] https://metr.org/assets/common_elements_of_frontier_ai_safety_policies.pdf

[10] https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[11] https://airisk.mit.edu/#:~:text=The%20Causal%20Taxonomy%20of%20AI,False%20or%20misleading%20information%E2%

[12] https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf

[13] https://arxiv.org/pdf/2307.03718

[14] https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international%5Fscientific%5Freport%5Fon%5Fthe%5Fsafety%5Fof%5Fadvanced%5Fai%5Finterim%5Freport.pdf

[15] https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/668ed084aa1d110a6f850508/1720635525144/Coordinated+Disclosure.pdf

[16] https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf" \h

[17] https://researchbriefings.files.parliament.uk/documents/LLN-2024-0016/LLN-2024-0016.pdf

[18] https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international%5Fscientific%5Freport%5Fon%5Fthe%5Fsafety%5Fof%5Fadvanced%5Fai%5Finterim%5Freport.pdf

[19] https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update

[20] The value chain of general-purpose AI | Ada Lovelace Institute

[21] https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property

[22] https://www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/

[23] AI incident reporting: Addressing a gap in the UK's regulation of AI (longtermresilience.org)

[24] https://crfm.stanford.edu/open-fms/

[25] https://www.centerforcybersecuritypolicy.org/insights-and-research/ntia-report-reveals-support-for-open-ai-models#:~:text=As%20the%20report%20clearly%20indicates,model%20ecosystem%2C%20by%20collecting%20evide

[26] https://www.tandfonline.com/doi/full/10.1080/17579961.2023.2184135

[27] https://artificialintelligenceact.eu/article/3/#:~:text=%2863%29%20%27general%2Dpurpose,is%20placed%20on%20the%20market

[28] https://opensource.org/deepdive/drafts

[29] https://www.gov.uk/government/publications/government-cyber-security-strategy-2022-to-2030

[30] https://www.ncsc.gov.uk/

[31] https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer

[32] https://www.aisi.gov.uk/grants

[33] https://www.gov.uk/government/publications/public-bodies-handbook-part-21-setting-up-an-arms-length-body/setting-up-a-new-arms-length-body-alb-guidance-for-departments-html

[34] https://www.cnas.org/publications/commentary/regulating-artificial-intelligence-must-not-undermine-nists-integrity

[35] https://oms-www.files.svdcdn.com/production/downloads/academic/AISIs%20Roles%20in%20Governance%20Workshop.pdf?dm=1721117994

[36] https://www.aisi.gov.uk/work/safety-cases-at-aisi

[37] https://www.gov.uk/government/news/global-leaders-agree-to-launch-first-international-network-of-ai-safety-institutes-to-boost-understanding-of-ai

[38] https://www.oxfordmartin.ox.ac.uk/publications/aisis-roles-in-domestic-and-international-governance

[39] https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance

[40] https://www.science.org/doi/10.1126/science.adi8982

[41] https://www.ukri.org/

42 https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/

43 https://link.springer.com/article/10.1007/s43681-023-00289-2

44 https://www.aisi.gov.uk/grants

45 https://www.turing.ac.uk/news/publications/common-regulatory-capacity-ai

46 https://www.gov.uk/government/organisations/national-data-guardian

47 https://ico.org.uk/

48 Regulators' strategic approaches to AI - GOV.UK (www.gov.uk)

49 Large language models and generative AI (parliament.uk)

50 https://www.governance.ai/research-paper/open-problems-in-technical-ai-governance

51 https://crfm.stanford.edu/2023/11/18/tiers.html

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

53 https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations

54 https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

55 https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf

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

57 https://www.longtermresilience.org/post/transforming-risk-governance-at-frontier-ai-companies

58 https://www.nature.com/articles/s42256-021-00370-7

59 https://www.anthropic.com/news/model-safety-bug-bounty

60 https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice#:~:text=The%20Code%20of%20Practice%20will,of%20Practice%20to%20demonstrate%20compliance

61 https://crfm.stanford.edu/2023/11/18/tiers.html

62 https://artificialintelligenceact.eu/the-act/

63 https://www.governance.ai/research-paper/training-compute-thresholds-features-and-functions-in-ai-regulation

64 https://cohere.com/research/papers/on-the-limitations-of-compute-thresholds-as-a-governance-strategy-2024

65 https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill%5Fid=202320240SB1047

[66] https://arxiv.org/pdf/2407.05694v1

[67] https://commonslibrary.parliament.uk/research-briefings/cbp-10011/

[68] https://www.nature.com/articles/s41599-024-03017-1

[69] https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[70] https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024

[71] https://www.gov.uk/government/publications/international-recognition-procedure/international-recognition-procedure

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

[73] https://www.adalovelaceinstitute.org/report/under-the-radar

[74] https://arxiv.org/abs/2403.04893

[75] https://carnegieendowment.org/research/2024/03/charting-the-geopolitics-and-european-governance-of-artificial-intelligence?lang=en&center=europe

[76] https://facctconference.org/static/papers24/facct24-152.pdf

[77] Council of Europe opens first ever global treaty on AI for signature - Portal (coe.int)

[78] https://cetas.turing.ac.uk/publications/towards-secure-ai

[79] https://www.etsi.org/technologies/securing-artificial-intelligence

[80] https://aistandardshub.org/ai-standards/information-technology-artificial-intelligence-risk-management/

[81] https://www.rand.org/pubs/research%5Freports/RRA2849-1.html

[82] https://www.elysee.fr/en/ai-action-summit

[83] https://www.institute.global/insights/tech-and-digitalisation/exploring-eu-uk-collaboration-on-ai-a-strategic-agenda

[84] https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/

[85] https://labour.org.uk/change/mission-driven-government/

[86] https://www.gov.uk/government/publications/artificial-intelligence-ai-opportunities-action-plan-terms-of-reference/artificial-intelligence-ai-opportunities-action-plan-terms-of-reference

[87] https://arxiv.org/pdf/2307.03718

[88] https://arxiv.org/abs/2407.07300

TONY BLAIR
INSTITUTE FOR
GLOBAL CHANGE

Follow us

facebook.com/instituteglobal
twitter.com/instituteGC
instagram.com/institutegc

General enquiries

info@institute.global