THE CENTRE FOR
LONG-TERM RESILIENCE

Written evidence submitted to the Department for Science, Innovation and Technology: "Pro-Innovation Approach to AI Regulation" consultation

PREPARED BY

Dr. Jess Whittlestone and Nikhil Mulani

The Centre for Long-Term Resilience

THE CENTRE FOR
LONG-TERM RESILIENCE

CONTENTS

1. EXECUTIVE SUMMARY

The Centre for Long-Term Resilience (CLTR) is an independent think tank with a mission to transform global resilience to extreme risks. We do this by working with governments and other institutions to improve relevant governance, processes, and decision-making.

We were pleased to see the government's whitepaper on "A pro-innovation approach to AI regulation" published in March 2023. In many places, this whitepaper shows real promise for the UK to take a world-leading approach to AI regulation and effectively balance innovation and risk mitigation. We are particularly pleased to see the commitment to AI lifecycle accountability and the recognition that foundation models may need to be regulated differently from sector-specific applications.

As the government well recognises, a lot has happened in the world of AI since March. Major developments include heightened concern around societal-scale risks from AI, the UK government

announcing a commitment to global leadership in AI safety, and redoubled private sector investment into accelerating AI R&D (including the merger of Google Brain and DeepMind).

The pace of change is rapid and some parts of the whitepaper are at risk of quickly becoming outdated. (We mentioned that this could be a risk in our previous submission and engagement, but things have changed even more quickly than we expected (CLTR, 2022)).

In the wake of these new developments, our submission provides a risk-mapping framework that we think is helpful for defining the role of regulation and designing policies at specific intervention points. We then discuss some specific recommendations for the UK's approach to AI regulation based on this framework.

The framework examines how a range of risks from AI originate and can be mitigated by intervening at different points in the AI lifecycle. We highlight four main categories of risk:

1. **Risks from Model Development**: Emergence of dangerous capabilities, which could allow models to adopt goals different from those specified by the user.

2. **Risks from Proliferation**: If a model is widely accessible, and there are no protections in place for ensuring its capabilities are used safely, there is potential for misuse by malicious actors.

3. **Risks from Deployment**: Irresponsible or unsafe deployment in high-stakes domains could result in catastrophic accidents and damage to critical infrastructure.

4. **Risks from Wider Societal Impacts**: As AI systems are increasingly adopted, their longer-term consequences could include economic displacement, geopolitical tensions, and erosion of democratic systems.

We suggest that a successful regulatory approach should address three broad aims:

1. **Increasing Visibility:** The government needs visibility on how AI systems are developed and deployed, where sources of risk and gaps in risk management exist, and how any harms are felt by the public.

2. **Defining Best Practices:** Policymakers will need to translate their understanding of risks and harm into best practice guidelines for developers, users, and companies.

3. **Incentivising and Enforcing Best Practices**: Regulators can create incentives and penalties that encourage behavioural change by developers, companies, and users.

We use this framework to offer a high-level discussion of strengths and opportunities for improvement in the whitepaper's approach. We provide three detailed policy recommendations:

1. **Regulating Foundation Models**: Establishing technical standards and best practice guidelines required of all foundation model developers, deployers and users is necessary now to help begin preventing and mitigating risks. Additionally, laying the groundwork for frontier model regulation would begin reducing risks from dangerous capabilities and broader forms of misuse.

2. **Implementing an Information-Sharing Pilot**: Setting up a voluntary information-sharing pilot program with leading AI labs, centred on model capability evaluations and compute usage, would provide the government with the visibility necessary for designing well-informed regulation and risk functions.

3. **Best Practices for a Central Risk Function**: We endorse the whitepaper's proposed creation of a central function for AI monitoring and evaluation, and we recommend that the government apply risk management best practices from the "three lines of defence" model to the design and operations of this function.

2. BACKGROUND: SOURCES OF AI RISK AND THE ROLE OF REGULATION

In this section, we briefly outline a framework for thinking about different risks posed by AI and the role of regulation as a tool for managing those risks.[1]

There is no perfect way to categorise risks from AI - different categorisations will foreground different aspects of risk, and be useful for different purposes. For the purpose of identifying policy levers and intervention points, we suggest that it may be particularly useful to delineate how risks arise at different points in the AI lifecycle, from the initial development of a system through to its adoption in specific applications and wider impact on society.

[1] The framework summary throughout this section is adapted from "AI Risk: Advancing the Policy Debate," a forthcoming report jointly authored by the Centre for Long-Term Resilience and the Centre for Emerging Technology and Security. We recommend consulting this report for a more detailed treatment of the framework and its implications for domestic and international UK AI policy.

With this in mind, we suggest considering risks that arise at four important stages of the AI lifecycle: risks arising at the point of initial model *development*; risks that arise when access to AI capabilities *proliferate* more widely; risks arising from the *deployment* of AI capabilities in specific domains; and risks arising from the wider *societal impacts* of AI applications.

1. **Risks from Model Development**: the emergence of dangerous capabilities (Perez et. al. 2022; Shevlane et. al. 2023).

   The development of increasingly capable foundation models has been shown to lead to the emergence of hard-to-predict, often concerning capabilities, including "sycophancy" (repeating a user's preferences back to them), expressed desires for resource acquisition and goal preservation, and the expression of more extreme political viewpoints as models undergo further reinforcement learning from human feedback (Perez et. al. 2022).

   There is increasing concern that ongoing AI development could lead to even more dangerous capabilities, including models that pursue long-term, real-world goals different from the ones specified by users (Chan et al., 2023; Ngo et al., 2022), and power-seeking behaviours (such as avoiding shutdown, even when requested by users) (Krakovna and Kramar, 2023).

2. **Risks from Proliferation**: potential for misuse (Brundage et. al. 2018).

   Once highly capable AI models become widely available, its capabilities can be used for purposes not intended by the developer. If guardrails such as user verification or capability restrictions are not in place, otherwise beneficial capabilities of AI models can be repurposed for ill intentions (known as "dual-use" potential).

   Examples of "dual-use" capabilities of particular concern could include repurposing a model for offensive cyber capabilities, creating or acquiring destructive weaponry, or exerting control and surveillance over populations (Brundage et al, 2018).

3. **Risks from Deployment**: irresponsible or unsafe deployment in high-stakes domains (Arnold and Toner 2021).

Risks from deployment arise at the point where an AI system is applied in a specific domain. If proper checks and testing are not in place, AI accidents could include critical infrastructure failures such as electricity grid blackouts, or risky behaviour in sensitive domains such as unintentional conflict escalation (Arnold and Toner, 2021). Accidents could also stem from the previously mentioned emergent dangerous capabilities, including deception and power-seeking behaviours by misaligned AI systems (Ngo et al, 2022).

4. **Risks from Wider Societal Impacts**: including economic displacement, geopolitical tensions, or erosion of democratic systems.

   Risks from structural impacts arise more diffusely, from the gradual changes that continued adoption of advanced AI systems could foster throughout society. Structural impacts could include economic impacts (Furman and Seaman, 2019), geopolitical instability (Ding and Dafoe, 2021), exacerbated inequalities and discrimination (Babuta and Oswald, 2019), as well as reduced information quality and limitations on freedom of thought (Seger et al, 2020).

An effective regulatory approach should include mechanisms to anticipate and address the full range of risks, but should also target sources of risk as carefully as possible so as to avoid overreach and not unduly constrain beneficial innovation.

This categorisation of risks can then help us to think about the kinds of policy interventions needed to address each in a targeted way:

- Risks arising primarily at the point of model development, such as the emergence of dangerous capabilities, suggest we need **oversight and guardrails over the development processes** of some models, particularly those pushing the frontier of general-purpose capabilities. In particular, we need better regulatory oversight of the forefront of AI development, and better methods to evaluate and scrutinise models for potentially dangerous capabilities. We may also need to consider more strict limits on development, such as compute caps on training runs or moratoria on certain kinds of research, if societal risks seem sufficiently large. We discuss these policy options in more detail when discussing the regulation of foundation models in the next section.

- Risks arising at the point of model proliferation, particularly misuse risks, suggest we may need some **regulatory restrictions on widespread access to certain systems and capabilities**, informed by risk assessments of capabilities at the development stage. Policy

levers to consider here include guidelines for and possibly restrictions on open-sourcing, requirements for companies to have know-your-customer (KYC) processes for certain models, and potentially export controls.

- To address risks arising from irresponsible or unsafe deployment of AI systems, we need **oversight and guardrails around the deployment of AI in certain high-stakes domains**, including clear standards and assurance processes, and potential exclusion zones (contexts where AI should not be deployed, e.g., possibly in nuclear command and control).

- Addressing risks from the wider societal impacts of AI development and applications is perhaps the most challenging, as risks will arise gradually over time with no clear intervention point. However, the government can invest in **better methods for anticipating these risks and identifying early warning signs** (e.g., paying attention to early evidence of harms on smaller scales), and more diverse involvement in the development, deployment and scrutiny of models.

Policy Levers Across AI Lifecycle[2]

	POLICY LEVERS		
	CREATING VISIBILITY	**DEFINING BEST PRACTICES**	**INCENTIVES & REGULATION**
DEVELOPMENT AND TRAINING	• MODEL REPORTING AND INFORMATION-SHARING • THIRD-PARTY AUDITING ECOSYSTEM	• ORGANISATIONAL GOVERNANCE AND RISK MANAGEMENT GUIDELINES • MODEL DESIGN STANDARDS • PRIVACY-PRESERVING TRAINING AND AUDITS	• AI ASSURANCE ECOSYSTEM • PUBLIC R&D FUNDING ECOSYSTEM • LICENSING DEVELOPERS
MODEL PROLIFERATION	• OSINT TRACKING OF MODEL PROLIFERATION • WATERMARKING & AUTHORSHIP VERIFICATION	• OPEN-SOURCING GUIDELINES • KNOW-YOUR-CUSTOMER PRACTICES	• LICENSING USAGE • EXPORT CONTROLS
RESPONSIBLE DEPLOYMENT	• INCIDENT SHARING • AI BOUNTIES	• SENSITIVE APPLICATION AREAS AND NOVEL CHARACTERISTICS • POST-DEPLOYMENT MONITORING	• LEGAL LIABILITIES • CERTIFICATION
STRUCTURAL IMPACTS	• MEASURING JOB DISPLACEMENT • EVALUATION OF AI INNOVATION • UNDERSTANDING PERCEPTIONS OF AI PACE OF CHANGE	• INFORMATION QUALITY • ANTI-TRUST AND SAFETY COOPERATION • EMPOWERING UNDER-REPRESENTED GROUPS IN GOVERNANCE • DEVELOPING SKILLS TO ADDRESS AI IMPACTS	• INVESTMENT SCREENING • PUBLIC COMPUTE RESOURCES • REDISTRIBUTIVE ECONOMIC POLICIES

One of the challenges to be addressed here is clearly delineating the scope of policy intervention for each of these types of risk. This is a question the government will want to seek expert input

[2] Adapted from "AI Risk: Advancing the Policy Debate," a forthcoming report jointly authored by the Centre for Long-Term Resilience and the Centre for Emerging Technology and Security

and advice on, and will want to be able to navigate in a flexible manner as the boundaries of what poses a risk continue to change.

Across all sources of risk, we believe that a regulatory approach to AI needs to address three main policy goals:

- **Increasing Visibility:** The government needs visibility of how AI systems are developed and deployed, where sources of risk and gaps in risk management exist, and how any harms are felt by the public, including policies such as:
    - Model reporting and information-sharing
    - Incident sharing and whistleblowing systems
    - Measuring and evaluating harms and impacts

- **Defining Best Practices:** Policymakers will need to translate their understanding of risks and harm into best practice guidelines for developers, users, and companies, including:
    - Organisational governance and risk management guidelines for developers
    - Technical standards for model design
    - Third-party auditing
    - Guidelines and potential exclusion zones for high-risk areas
    - Empowering under-represented groups in governance discussions

- **Incentivising and Enforcing Best Practices**: Regulators can create incentives and penalties that encourage behavioural change by developers, companies, and users, including by:
    - Building an assurance ecosystem
    - Licensing developers or usage
    - Liability
    - Investment screening
    - Public compute resources

In general, the level of stringency needed across visibility, best practices, and regulation should depend on the assessed level of risk. In this sense, AI regulation definitely should be proportionate and risk-sensitive. This does not necessarily mean that AI regulation needs to be context-sensitive (in the sense of being scoped to specific applications or industries), but rather that it will be effective if it is intentionally designed towards reducing specific types of risks.

THE CENTRE FOR
LONG-TERM RESILIENCE

3. THE UK'S REGULATORY APPROACH: STRENGTHS & OPPORTUNITIES FOR IMPROVEMENT

The regulatory strategy outlined in the "Pro-innovation approach" whitepaper - particularly its assessment of lifecycle accountability as critical to effective AI regulation, its focus on foundation models as an important area for further attention, and its consideration of a central function for horizon-scanning and cross-sectoral risk assessment - show a promising start to the U.K.'s efforts to build out a forward-looking regulatory strategy that solidifies the country's leadership in AI.

However, we also believe that the current regulatory approach over-indexes on risks arising from sector-specific applications relative to the other categories of risk delineated in the previous section. The whitepaper states that "[The government] will not assign rules or risk levels to entire sectors or technologies. Instead, we will regulate based on the outcomes AI is likely to generate in particular applications." This approach will cover risks from the deployment of AI in specific domains and some (but not all) potential for misuse. By mostly focusing on intervening at one point in the AI lifecycle - the point where an AI system is deployed in a specific application - this context-specific approach will fail to adequately address the full range of AI risks, in particular:

- **This sector-specific approach will not address risks from dangerous capabilities** (such as emergent power-seeking capabilities) **and broader forms of misuse** (such as generating and disseminating disinformation), which could originate from frontier AI research and development

- **Nor will it address the more diffuse structural effects that develop over time** (including job displacement and increased geopolitical tensions), as AI systems scale and are used more widely throughout society

To start making headway on these risks, much greater visibility over AI development, deployment and impacts - especially at the frontier of development and concerning wider societal impacts - is needed. The whitepaper acknowledges that some visibility into coming developments will be necessary to ensure that the AI regulatory framework remains effective. In particular, the whitepaper proposes government functions whose responsibilities encompass cross-sectoral risk monitoring and horizon-scanning for emerging risks and opportunities. We propose these efforts be emphasized and focused towards high-risk frontier AI systems as well as on anticipating the more diffuse impacts of AI systems on society over time. Once greater visibility of frontier

development is established, the resulting increased understanding of sources of risk should be used to inform the design and implementation of regulations that reduce risks from new research and development.

We believe there are ways to design AI policies that are effective at reducing the largest risks while still being proportionate and carefully targeted. In particular, any regulation focused on the development of AI systems should be focused on a small set of frontier AI systems rather than on specific applications.

The whitepaper claims it would be "premature to take specific regulatory action in response to foundation models." However, we would argue that the recent rapid developments in AI suggest that now is an ideal time to begin regulating, in order to get ahead of reducing the risks from dangerous capabilities and misuse that frontier models could pose. **While believing there is a pressing need to begin taking regulatory action, we recommend that any policy action be careful and incremental.** The government should take an iterative approach that leaves room for flexibility as we continue to observe and learn more about risks and opportunities from frontier AI development. We provide more detailed recommendations on what steps can be taken to regulate foundation models now, and where more thinking is needed, in recommendation 1 below.

RECOMMENDATION 1: REGULATING FOUNDATION MODELS

Our top recommendation is that the government should, with some urgency, lay out a clear approach to regulating foundation models - with a particular focus on foundation models at the 'frontier' of AI development, as particularly high-risk.[3]

However, we also recommend that foundation model regulation should be careful, incremental, and clearly motivated by specific risks (such as those laid out in the background section of this paper), so as not to unduly limit innovation.

As the whitepaper acknowledges, foundation models are particularly deserving of regulatory attention for two reasons: **(1) due to their complex lifecycle and the need to distribute accountability across that lifecycle,** and **(2) because they are known to produce emergent, hard-to-predict capabilities which could pose extreme risks.** First, the development and

[3] On frontier AI regulation, see also: Anderljung, Barnhart, Leung, Korinek, O'Keefe, Whittlestone, et al (2023). "Frontier AI Regulation: Managing Emerging Risks to Public Safety" [*Forthcoming*]

deployment lifecycle for foundation models, in which some actors build general-purpose foundation models and others customize and apply these models in specific domains, means resulting risks are complex to manage and require attention from both upstream developers and downstream deployers. Second, as outlined in the framework above, the emergence of novel capabilities in foundation models introduces risks from dangerous capabilities and a wide array of forms of misuse, suggesting more stringent oversight is needed.

These two motivations for regulating foundation models can easily be confused, but suggest slightly different regulatory needs.

The challenge of addressing life-cycle accountability suggests that all foundation model developers should be expected to follow certain standards and best practices for safe and responsible development, including, for example, requirements to conduct risk assessments prior to and throughout training (Shevlane et al, 2023), reporting and documentation of model characteristics (Mitchell et al 2018), and external auditing of model design and development (Mökander et al, 2023). The full range of standards required of all foundation model developers should be assessed through consultation with a wide set of civil society and technical stakeholders throughout academia, the private sector, non-profits, and the general public. Establishing some standards for safe and responsible foundation model development would help address the full range of risks from AI, by capturing potential sources of risk at the development stage and ensuring they are not propagated or exacerbated once systems proliferate, are deployed in specific sectors, and have wider impacts on society (though of course will also need to be complemented by additional interventions at each of these later stages).

In addition to requirements for all foundation model developers, we believe that **additional, stricter, requirements should be placed on those actors developing frontier AI systems** - those developing the “cutting edge” systems, where particularly dangerous capabilities could emerge unexpectedly. This might include requirements to report all training runs above a certain compute threshold to a regulatory authority, to demonstrate to that authority that certain evaluations and safety tests have been conducted before releasing a system, and to adapt release strategies to the level of risk assessed in the system.

Precisely defining what counts as a “frontier” foundation model is not entirely straightforward, as there are many different ways to push the frontier of capabilities. For now, we suggest using a training compute threshold such as 1e26 FLOP, since training compute empirically correlates with breadth and depth of capabilities in foundation models (Owen, 2023; Sevilla et al, 2022). This could be supplemented over time with additional measures which look at whether a model

exceeds the capabilities of the most capable models already available and known to not have dangerous capabilities.

Stricter enforcement mechanisms for standards around frontier AI development, and greater regulatory visibility over such development, will also likely be needed. Policies for enforcement could include licensing for model developers, bans on certain types of dangerous research, legal liability for harms caused by models, and mandatory periodic auditing by public or third-party bodies. Such methods of enforcement (especially licensing and mandatory reporting or auditing requirements) should be carefully designed in order to avoid inadvertently further entrenching existing AI leaders and creating barriers to entry.

Referring back to our earlier framework of policy aims, we suggest the following concrete steps that the government could take in regulating foundation models:

- Increasing Visibility:
    - **Pilot an information-sharing program** which allows foundation model developers to voluntarily share information about model development and characteristics with a government authority - more details in recommendation 2 below;
    - **Explore options and thresholds for a mandatory reporting program**, which could build on the lessons of the pilot, and require developers to report training runs above a certain threshold.
- Defining Best Practices:
    - **Convene diverse expertise across industry, academia, and civil society, to develop a set of standards and best practices** for both (a) foundation model development in general, and (b) frontier foundation model development in particular.
- Incentivising and Enforcing Best Practices:
    - **Explore initial, light-touch ways to incentivise adherence** to best practices for all foundation model developers, such as a third-party auditing and certification scheme
    - **Explore more stringent enforcement mechanisms** for frontier AI development and their pros and cons, including licensing, liability, and mandatory periodic auditing.

RECOMMENDATION 2: IMPLEMENTING AN INFORMATION-SHARING PILOT

THE CENTRE FOR
LONG-TERM RESILIENCE

We recommend implementing a structured information-sharing program to create visibility around frontier AI development activities and inform foundation model regulation.

It is heartening that the government has recently made progress towards greater frontier visibility by securing pledges for early model access from leading AI labs. We suggest the government build upon this commitment by setting up a voluntary information-sharing pilot program with leading AI labs centred on foundation model capability evaluations and compute usage (Mulani and Whittlestone, 2023).[4]

The Office for Artificial Intelligence seems a natural home for such information-sharing, but it could also potentially be structured as an initiative within the Foundation Models Taskforce or as a more independent body akin to the Cyber Security Information Sharing Partnership within the National Cyber Security Centre.

An information-sharing pilot could provide visibility useful both for informing targeted foundation model regulation as well as standing up a central risk function, through creating the following processes:

1. Training run reports containing model capability evaluations and compute usage could be shared by labs to the government for a limited subset of new foundation models, which are especially compute-intensive or have especially general capabilities[5]

2. Continually updated versions of the training run report could be shared before and throughout model training and deployment processes

3. Shortly before release, labs could also grant the Office direct access to their models

4. Subject-matter experts employed by the Office – or seconded from partner organisations – could analyse the information provided by labs for its implications about current risks and forthcoming developments

5. These experts could provide actionable policy recommendations to relevant stakeholders throughout the government. For instance, recommendations could inform technical and governance standards, as well as compute or capability benchmarks for regulation applied to frontier models.

[4] This recommendation summarizes a detailed proposal for implementation of an information-sharing pilot, which is available here
[5] See Mulani and Whittlestone, 2023 for more details on what training run reports could look like

THE CENTRE FOR
LONG-TERM RESILIENCE

RECOMMENDATION 3: BEST PRACTICES FOR A CENTRAL RISK FUNCTION

We endorse the whitepaper's proposed creation of a central function for AI monitoring and evaluation, and we recommend that the government apply risk management best practices from the "three lines of defence" model to the design and operations of this function.[6]

The "three lines of defence" model provides a structure for ensuring that risk management activities include capacities for accountability from a governing body, execution towards risk management goals, and independent assurance and advice to promote rigour (IIA, 2020). If implemented according to the "three lines of defence model," the central function could help ensure that the AI regulation framework is implemented, enforced, and updated on an ongoing basis in a robust and sensible manner.

In order to be effective, the central function should be designated as the owner for AI-related risk matters across the government. The function should be accountable for all AI risk, but responsibility for different aspects of AI risk could sit across multiple departments. It would seem sensible for this leadership to sit within DSIT, but our recommendation is primarily for a central point of ownership accountability rather than for a particular departmental home. In practice, such accountability would include coordinating the identification, assessment, prevention, mitigation, and reporting of AI-related risk across government.[7] Our prior recommendations could help with each of these tasks: information-sharing would be valuable for risk identification and assessment, while foundation model regulation would be valuable for risk prevention and mitigation.

Central ownership of AI risk would constitute the first line of a "three lines of defence" model for risk management across the government, alongside units in other departments that take ownership over other risks.

AI risk ownership should be overseen and supported by an Office of Risk Management (ideally headed by a government Chief Risk Officer) within the Cabinet Office, constituting the second line of defence. This Office would be responsible for risk oversight and the overall risk management process across the government.

[6] See CLTR's 2022 report on extreme risks and the UK National Resilience Strategy for additional discussion of the three lines of defence model

[7] This proposed central unit that owns accountability for AI risks is similar in intention to the newly announced Central Biological Security Coordination Unit (although this Biological unit sits in the Cabinet Office)

In turn, this risk oversight would be scrutinised by an independent National Resilience Institute, comprised of external experts from the AI field and other relevant domains. This Institute would form the third line of defence, accountable to Parliament. Industry regulators would oversee and audit AI developers directly and could report to the National Resilience Institute.

Corporate developers should be required to adopt risk management best practices within their governance structures and encouraged to adopt their own 'three lines of defence' risk management frameworks (Schuett, 2022). This would include designating their own Chief Risk Officers, as well as internal and external audit teams responsible for overseeing risk ownership and reporting to the board of directors.