Written Evidence Submitted by Dr Jess Whittlestone and Richard Moulange
(GAI0071)

Prepared by:

Dr Jess Whittlestone — Head of AI Policy, Centre for Long-Term Resilience
Richard Moulange — PhD student, MRC Biostatistics Unit, University of Cambridge

Introduction

1. 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.

2. We are pleased to see the Committee's inquiry into the governance of artificial intelligence in the UK. While AI has potential to bring many benefits, we are concerned that extreme risks could also arise from AI development, particularly given rapid progress in AI capabilities in recent years. As AI capabilities are integrated into increasingly high-stakes areas such as defence and critical infrastructure, mistakes could be catastrophic; AI-related power shifts could threaten international strategic stability; and the development of powerful advanced AI systems could leave humanity with little or no control over the future (Clarke and Whittlestone 2022). Building on the UK's unique position as home to both pioneering AI companies and world-class expertise in AI risk, and existing commitments to prioritise safe and responsible AI development (such as those in the National AI Strategy and Defence AI Strategy), **the UK is well-placed to become a world leader in understanding and managing extreme risks from AI**.

3. This submission makes three key recommendations for strengthening the UK's approach to AI governance, along with some questions we suggest the Committee asks the Government about its current governance plans.

Recommendation A: Ensure that the UK's regulatory regime is supported by a strong broader governance ecosystem which can provide expertise and identify and address regulatory gaps.

4. Current governance of AI in the UK is fragmented and dependent on regulation in related areas (such as data protection) which is not designed specifically for AI. The National AI Strategy and policy paper 'Establishing a pro-innovation approach to regulating AI' recognise the limitations of the existing landscape, and the need for regulation more specifically tailored to the challenges and risks posed by AI.

5. The regulatory approach laid out in the policy paper emphasises supporting existing regulators to identify and address challenges posed by AI in their domains. We agree that the governance of AI will often need to be tailored to the context of application, and that the UK should begin by drawing and building on the strengths of existing regulators.

6. However, we also believe this context-specific approach faces several important challenges, including ensuring regulators have sufficient expertise and capacity to understand the implications of AI; ensuring that regulatory gaps can be identified and addressed; and being sufficiently adaptive to advances in AI capabilities.

7. For the regulatory approach to successfully navigate these challenges, we believe it is crucial that the regulatory regime is **supported by a broader governance ecosystem** which can effectively identify and address inefficiencies and gaps. In practice, this means identifying or creating a body (or bodies) with a clear mandate and the capacity to:

a. **provide AI-related expertise and upskilling for regulators**
b. **identify cross-sector lessons and promote coherence**
c. **identify and anticipate regulatory gaps and ensure they are acted upon**

8. We do not think that there is an already-existing body which has sufficient scope, capacity, and power to fill all these gaps. For example, though an independent body like the Alan Turing Institute may be well-suited to acting as a central hub of AI expertise for AI regulators, identifying and acting on regulatory gaps seems better positioned within a central government body like the CDEI with a closer relationship to regulators. We therefore believe that the government should either establish a new body to take on these responsibilities, or explore how a broader governance ecosystem, composed of several different actors with different responsibilities and access to policy levers, can provide regulators with the resources and information they need to regulate AI effectively. Since creating another body may increase the complexity of the regulatory regime further, it may be better for the Government to first explore how to empower current organisations (such as the CDEI and Office for AI) with clearer mandates and, crucially, specific statutory authority.

9. The Committee may wish to ask:
   a. How does the Government propose to anticipate, identify, and address regulatory gaps – for example, issues that do not fall neatly within the remit of existing regulators?
   b. How will the Government ensure that regulators have access to sufficient and appropriate expertise to understand and anticipate the impacts of AI on their domain?
   c. How can the Government empower existing bodies such as the CDEI and Office for AI to better support regulators?

Recommendation B: Consider that more cross-cutting and anticipatory regulation may be needed to mitigate extreme risks, especially as capabilities advance.

10. While we support the current government focus on creating "proportionate [and] light-touch" regulation, and avoiding unnecessary barriers to innovation, we would like to see a stronger presumption that some types of AI systems and applications will require additional legislation and cross-regulatory attention. By developing a clear risk framework and robust ways to identify particularly high-risk areas of AI development and deployment, the Government can effectively mitigate extreme risks while leaving beneficial innovation unencumbered.

11. This is especially important given the rapid pace of AI progress. As advances in AI research lead to increasingly general capabilities with increasingly wide-ranging societal ramifications, more anticipatory and cross-cutting regulatory powers will be needed for the Government to identify and address potential risks.

12. In particular, **access to increasingly large amounts of computing power ('compute') has recently enabled dramatic progress in 'foundation models'** – models trained on broad data that can be adapted to a range of downstream tasks, including language and image generation models (Kaplan et al. 2020; Sevilla et al. 2022; Bommassani et al. 2022). Foundation models are showing impressively general performance across a wide range of tasks, unlocking many capabilities in the past year which experts expected to take 5 to 10 years – including photorealistic image generation, realistic text to video generation, and code generation. Governments around the world are increasingly getting caught off-guard by the policy implications of these developments, such as harmful biases displayed by computer vision and natural language processing systems. There are good reasons to expect this trend of compute-intensive AI development to continue, leading to increasingly general-purpose AI systems from which a large proportion of societal applications derive (Kaplan et al.

2020). This means that, in the future, the Government is even more likely to face significant AI governance challenges that it cannot easily predict.

13. These trends in AI progress could challenge the UK's existing approach to AI regulation in two important ways:

a. **It is increasingly important to distribute responsibility across the entire supply chain of AI development**. If societal applications of AI increasingly come from downstream applications of foundation models, only regulating at the application level is likely to put an increasing and undue burden on SMEs relative to the developers of foundation models. The regulatory approach must consider at what stage of the supply chain potential risks or harms are most easily detected, and who is best placed to take action to mitigate the risk (Engler and Renda 2022).

b. **There is a growing need for anticipatory governance approaches in addition to addressing specific harms as they arise on a sector-by-sector basis**. Progress towards more generally capable AI systems is likely to lead to more systemic, high-stakes and difficult to anticipate impacts on society, making a sector-specific approach to regulation increasingly challenging. For example, the use of increasingly advanced AI systems by militaries could threaten international stability or lead to unintentional escalation; increasingly capable AI systems used for content generation and persuasion could undermine democratic debate; and the application of AI to dual-use scientific research such as biotechnology could make it easier for rogue actors to develop or use dangerous technologies (Clarke and Whittlestone 2022). The UK government must therefore consider how to proactively shape AI innovation in ways that prevent harms from occurring.

14. To address these challenges, we suggest that the UK should:

a. **Recognise that some form of regulation may be needed for general-purpose systems such as foundation models in future and** commit to exploring the details and feasibility of such regulation further. There are many ways that developers of foundation models (or other actors further up the supply chain) might be held appropriately accountable for harms from AI development without hindering innovation. For example:

   i. We might **require developers of foundation models to cooperate with regulators** to identify and prevent cases of potential misuse

   ii. We might require developers of foundation models to **follow and document certain best practices** to ensure that systems function reliably and robustly and make use of sufficiently representative datasets. This could reduce the regulatory burden on smaller businesses: a business

using a foundation model which has already met certain requirements might then be subject to fewer requirements on the application level.

b. **Explore ways to collect better information about where potential harms might occur from AI development, enabling earlier intervention**.
    i. One way to do this would be to **collect and monitor data about compute usage**. Since compute-intensive AI systems are particularly likely to precipitate unexpected harms, tracking where and how large amounts of compute are being used can provide valuable information about where to focus anticipatory governance and risk assessment efforts. This could begin with the Government collecting data on broad compute trends, leveraging information provided from financial reporting, import duties, export controls, and information volunteered by AI companies and researchers in government-run foresight exercises. Eventually, the government could create reporting requirements for both users of compute (AI developers) and providers of compute (data centre operators), with reporting thresholds negotiated to capture systems of interest while minimising burden on low-risk innovation.
    ii. Beginning to collect and act on information about compute usage could strongly support the government's aim of establishing pro-innovation approach to regulating AI: by **making it easier in future to systematically identify potentially high-risk capabilities ahead of time, the government can more effectively direct regulatory attention and risk-assessment to those capabilities**, while leaving innovation in low-risk areas relatively unencumbered. See our submission to the Future of Compute Review for more details on this proposal.

15. The Committee may wish to ask:
    a. How will the Government ensure it has the information it needs to identify high-risk AI systems before deployment?
    b. How will the Government ensure that responsibility for the safe and responsible use of AI is distributed along the entire supply chain, including developers of general-purpose or foundation models?
    c. How will the Government prepare for the possibility that more cross-cutting regulation may be needed as AI capabilities advance? What kinds of new regulatory powers might be needed?

Recommendation C: Engage internationally to mitigate shared risks from AI

16. Many of the extreme risks posed by AI are not just risks to the UK, but to international prosperity and stability more broadly. Even if the UK develops AI safely and responsibly, we may face risks from the development of AI in other parts of the world. This means that the UK must engage internationally on issues of AI governance, starting by sharing best practices as an example for other countries, and encouraging joint commitments with other nations.

17. This is particularly important where developments in AI could pose threats to national and international security. Concerns include:

a. The misuse of AI by authoritarian states to **suppress dissent or subvert democratic processes** (Brundage, Avin et al. 2018), or by malicious actors to **develop novel biological or chemical weapons** (Urbina et al 2022)
b. Competitive pressures to field new AI capabilities ahead of adversaries, despite the fact that **modern AI systems often still fail or misbehave in unpredictable and inexplicable ways** (Horowitz and Scharre 2021)
c. **Changes to the character of warfare increasing the risk of unintended escalation**: the introduction of AI into military operations is likely to accelerate the tempo of warfare, which could result in scenarios where minor tactical missteps lead to a situation escalating out of control before any human has time to intervene, or where it is difficult to terminate or step back from conflict even when desired (Horowitz and Scharre 2021)
d. The use of AI **undermining nuclear deterrence** by making it easier for states to discover and destroy previously secure nuclear launch facilities (Lieber and Press 2017)

18. We are pleased to see many important commitments in the UK Defence AI Strategy: to the safe and responsible use of AI and to shape global AI developments to promote security, stability, and democratic values. We would strongly recommend that this Committee support the MOD, and wider national security community in government, to:

a. **Adapt existing safety and assurance processes** to the challenges raised by AI and develop governance processes to **ensure AI is deployed at the appropriate pace** given the capabilities and limitations of AI systems.
b. **Establish international dialogue to reduce risks from competition and misunderstanding**, including by:
    i. sharing best practices for the safe and responsible development of military AI systems (e.g., publicly communicating about the importance of robust test, evaluation, and assurance processes)
    ii. engaging in international dialogue to establish “rules of the road” for military AI use, including questions of what degree of automation is acceptable in what contexts; how to prevent the speed of military decision-

making from outpacing human control; and limits on the use of AI in nuclear command and control systems.

c. **Explore ways to prevent misuse of high-risk applications of AI**, for example by:
    i. Establishing collaborative efforts to better understand what kinds of AI systems are being misused and where. This could focus on whether such systems are being used by non-state actors to spread disinformation, to radicalise or to develop capabilities to carry out attacks; or by rival states to develop military capabilities, interfere in elections, or to violate human rights domestically
    ii. Consider ways to monitor and potentially restrict the use of AI models that may pose a significant threat to (inter)national security, such as via reporting requirements (as discussed in Recommendation 2), Know-Your-Customer (KYC) for developers and cloud providers, or export controls

(November 2022)