Written Evidence Submitted by DeepMind
(GAI0100)

Contents

DeepMind welcomes the opportunity to provide evidence to this important inquiry into the *governance of artificial intelligence.* AI can deliver **transformative benefits** for the UK and the world - by accelerating scientific discovery, helping to address grand challenges like climate change, and boosting innovation, productivity growth and standards of living. However, it can also create **risks** - and so it must be built, used and governed responsibly. The UK is a leader in AI, and has an opportunity to demonstrate to the world how this technology can be governed effectively in a way which supports innovation whilst mitigating social risks, and to act as a key broker in driving international coordination.

About DeepMind

DeepMind's mission is to solve intelligence to advance science and benefit humanity. At the heart of our work is a commitment to act as responsible pioneers in the field of AI, in service of society's needs and expectations.

The company's breakthroughs include AlphaGo, AlphaFold, more than one thousand published research papers (including over twenty in *Nature* and *Science*), partnerships with scientific organisations, and hundreds of contributions to Google's products (in everything from Android battery efficiency to Assistant text-to-speech). We were acquired by Google in 2014, but continue to be headquartered in the UK.

We know that putting our mission of solving intelligence to advance science and benefit humanity into practice comes with responsibilities. We have a set of clear Operating Principles to guide these efforts - defining both our commitment to prioritising widespread benefit, as well as the areas of research and applications we refuse to pursue. These principles have been at the heart of our decision making since DeepMind was founded, and continue to be refined as the AI landscape changes.

We invest in safety, ethics and sociotechnical research to inform our own approach to AI development and governance - and to help foster progress across the field of AI. For example, we recently assessed the near and longer-term ethics and safety risks posed by language models - a priority area of focus for AI researchers. We're also committed to contributing to industry best practices and standards on safety and ethics, such as our recent work with the Partnership on AI to develop best practices and processes for responsible human data collection.

Question 1.

How effective is current governance of AI in the UK? What are the current strengths and weaknesses of current arrangements, including for research?

- **AI is currently governed in the UK through a range of different approaches.** Regulation is one core element - and the UK has taken a decentralised approach to date, with individual regulators taking the lead developing responses in their domains.
- Notable examples of regulators taking a proactive approach to the use of AI in their domains include the ICO's work on data protection rights and AI, and joint work between the FCA and the Bank of England on ML in financial services.
- DeepMind believes this context-based approach, led by individual regulators, has some important advantages - helping to ensure regulation is informed by deep domain expertise; tailored to the dynamics and complexity within AI supply chains; proportionate to the risk of harm and sensitive to the context of deployment; and built on the strengths of existing regulatory approaches. However, it will also require continued investment and careful design to avoid becoming fragmented or uncoordinated over time. We discuss these strengths and considerations in more detail in our response to Question 4.
- Wider elements of the UK's AI governance ecosystem, beyond regulation, include: academic and industry R&D on core governance challenges like interpretability, safety and privacy (both fundamental research and the development of concrete tools and practices); forums for developers, users and those affected by AI systems to debate and agree best practices and standards; and early efforts from the Centre for Data Ethics and Innovation (CDEI) to encourage a robust assurance ecosystem to provides services that ensure AI systems work as they are supposed to. The UK also has a healthy ecosystem of research organisations such as the Ada Lovelace Institute, looking at the wider social implications of AI.
- This portfolio of AI governance approaches is steadily maturing. There are, however, at least three important challenges the UK will need to meet.

1. **More evidence on the scope and effectiveness of existing regulation that applies to AI**. As the Government's recent Policy Statement on Establishing a Pro-innovation Approach to Regulating AI noted, *"there are no UK laws that were explicitly written to regulate AI, it is partially regulated through a patchwork of legal and regulatory requirements built for other purposes".*

   It's important, therefore, that the forthcoming White Paper is based on a thorough assessment of the effectiveness of current law and regulation at managing risks from AI - and also identifies any barriers to innovation. If we simply adapt existing regulation on a case-by-case basis, or in response to

specific incidents, the UK could end up with dated, complex or uncoordinated regulation that undermines innovation and deters investment. We welcome the direction of travel in the Government's Policy Statement, but believe this should be supported by a gap analysis - conducted in partnership with key regulators - of existing law and regulation that applies to AI. This should be a collaborative and open process with opportunities for stakeholder feedback.

2. **The need for joined-up thinking on the UK's AI governance ecosystem**. As it stands, there is a patchwork of coordination between the various components that make up the UK's AI governance ecosystem - whether it's regulation, standards, assurance or R&D. There is an opportunity for the Government's upcoming White Paper to address this and to knit together the growing but disconnected work across these areas into a coherent set of actions. Getting this right will be key to realising the ambition of pillar 3 of the UK's National AI Strategy: "*to build the most trusted and pro-innovation system for AI governance in the world*." It's also important that the AI governance ecosystem is embedded in every sector that uses AI.

3. **Sustained investment in the UK's AI governance ecosystem**. Despite many promising initiatives, overall the UK's governance ecosystem remains at an early stage of development. For example, we lack widely-agreed upon best practices and standards that AI products and services can be assessed against. This is also true elsewhere in the world and International Standards Development Organisations' work on AI is still at an early stage - but there is a real opportunity for the UK to take the lead. The new AI Standards Hub is a promising effort but will need proper investment and backing. Another important route forward is R&D investment - many desired attributes of AI systems, such as making them more robust, explainable and fair, are also open research problems; as is the broader question of how to best evaluate potential risks from AI. The acceleration in AI system capabilities and diffusion across important parts of the economy and society make the task more urgent. We need the same focus on innovation in ethics, safety and governance as we have seen on AI capability development.

DeepMind's approach to governance

Our Operating Principles have been at the heart of our decision making since DeepMind was founded, and continue to be refined as the AI landscape changes and grows. Through them, we commit to social benefit; scientific excellence and integrity; safety and ethics; accountability to people; sharing knowledge responsibly; and diversity, equity and inclusion. They also commit us to not pursuing harmful technologies; weapons; surveillance technology; and technologies whose purpose contravenes international law or human rights.

These written principles are only part of the puzzle – how they're put into practice is key. We've spent many years developing our own skills and processes for responsible governance, research and impact across DeepMind. To help empower our teams to pioneer responsibly and safeguard against harm, our interdisciplinary Institutional Review Committee (IRC) meets every two weeks to carefully evaluate DeepMind projects, papers, and collaborations.

We've carefully designed our review process to include rotating experts from a wide range of disciplines, with machine learning researchers, ethicists, and safety experts sitting alongside engineers, security experts, policy professionals, and more. These diverse voices regularly identify ways to expand the benefits of our technologies, suggest areas of research and applications to change or slow, and highlight projects where further external consultation is needed.

While we've made a lot of progress, many aspects of this lie in uncharted territory. We won't get it right every time and are committed to continual learning and iteration. That's why we recently published detailed reflections and lessons from one of our most complex and rewarding projects: establishing our approach to releasing AlphaFold.

Question 2.

What measures could make the use of AI more transparent and explainable to the public?

- **Together, improved transparency and explainability of AI - as well as interpretability, a closely related concept - can provide a range of benefits to both developers and users.** These qualities could enable increased model robustness, accuracy and understanding to help AI researchers develop safer systems. For users and the wider public, such efforts can help demonstrate how and why decisions were made - making them fairer and ultimately increasing both trust in AI systems and accountability from the labs developing them.
- Policymakers should focus on **increasing funding for R&D** on transparency and explainability, **evaluating and recommending assurance tools** or standards, and **supporting broader efforts to increase wider public understanding** of AI. Efforts are underway across industry and academia, focusing on 3 main areas:
    1. **Understanding how AI models work**
        - A key goal of explainability research is to provide researchers and developers with a *mechanistic* understanding of how AI systems work. For example, large language models can have billions of parameters and are often trained on large amounts of data. Although the architectures of these transformer models are designed by research engineers, they remain opaque, meaning it is not easy for humans, no matter how skilled, to easily understand how and why a specific algorithmic output is generated.
        - **Developing a deeper understanding of how and why these models work is an ongoing research priority for DeepMind, and many other AI labs**, as it could enable researchers to improve the performance of AI models and inform work to make them safer and more robust.
          Explainability research will also support efforts to apply AI in other domains, such as in scientific research, where domain specialists will often need to understand how an AI system is reaching its outputs.
        - There is a broad range of promising directions in explainability research, including studying which neurons in a deep neural network are activated at a given moment; analysing and 'probing' the behaviour of AI agents across different types of data inputs tasking language agents with explaining their outputs to users, including by citing sources; and using approaches to AI such as 'program synthesis', that are more interpretable by default.
        - To support this work, the Government could fund and **incentivise greater research into explainability, and seek industry-wide agreement on priority research problems and evaluation methods to pursue**. This is

something which could be led in the UK by the AI Standards Hub and prioritised in UKRI's National AI R&D programme.

2. **Providing useful explanations to different audiences**
    - The challenge of providing useful explanations to end-users of AI-based applications, is different from the task of understanding AI systems from a mechanistic perspective. It requires a wider set of skills and should be informed by a rich understanding of the diverse needs of its users.
    - In certain contexts, people will quite reasonably want to know when they are interacting with an AI system. A common-sense approach would be to highlight to users, when relevant, proportionate and useful, that they are interacting with an AI system, or that AI was used as part of a decision. The original proposal from the European Commission for the AI Act included similar measures.
    - In cases where AI is being used to help make more consequential decisions, individuals will often want to know what variables or features a decision was based on, or what they can do to change the result. As highlighted by The Alan Turing Institute and the ICO in Project ExplAIn, useful explanations in such cases will always be context and user-specific. In many cases, partial explainability, which is tailored to the needs of the user and delivered in a useful manner, will be preferable – as many user questions can be reliably answered without fully understanding all aspects of how an AI system works.
    - There are a growing number of 'explainability' tools that can help explain such decisions to users. However, increasingly, the large number of such tools is becoming a challenge, as there is relatively little evaluation of them. This provides a clear role for government guidance, e.g. via a mature AI Assurance Ecosystem to vet and provide explainability tools.
    - Given these nuances, we welcome the Government's steer in its recent Policy Statement that regulators should be free to decide what *"transparency means for AI development or its use in the context of their sectors or domain*."

3. **Transparency on the use of AI**
    - The concept of 'transparency' in AI can be used to describe the efforts, summarised above, to make AI models more interpretable and explainable, to both practitioners and end-users. However, it can also be used much more broadly – to describe the activities that organisations and policymakers can invest in to provide information about AI systems, how they have been developed and evaluated, and how they are being used. Opportunities include:
    - **Documentation:** Organisations developing AI should provide clear

documentation for their models and datasets, such as model cards and data cards, to enable other organisations using these systems to do so in a safe and reliable way. DeepMind did this with our Gopher language model (appendix B), by setting out the intended uses, performance metrics and risks in plain english. Documentation like this can also provide information on how ethics and safety risks have been considered, as we recently did with our AlphaFold AI system. Governments can help by endorsing clear standards for these documentation practices, and could build on the model card approach, by setting out guidance as to what information regulators should expect in terms of proportionate transparency reporting.

- It's important to note that there will be limits to transparency, or situations where full release of information can pose disproportionate risks. For example, large generative models may pose risks in areas like privacy, and it might be right to limit access to such models behind an API (Application Programming Interface), so that experts can first study these risks in more detail, rather than fully open-sourcing the code.
- **Public AI literacy and sentiments:** Ultimately, all efforts to explain AI systems more effectively to the public need to be underpinned by a better understanding of public sentiments and knowledge about AI, and accompanied by broader public engagement efforts to communicate what AI systems are (and are not), and how they are being used across society.
- For example through our podcast, we aim to give the public the wider context about how AI is being developed and used at DeepMind, whilst through our work with the Raspberry Pi Foundation and other educational organisations, we aim to demystify AI and support the next generation of AI leaders.
- Government has a role to play here too, and work such as the CDEI AI Barometer, which measures trust and barriers to public understanding is a positive step towards an inclusive and informed conversation about the use of AI. Government could also consider taking inspiration from the Library of Congress' Digital Preservation initiative and archiving the datasets and models used in the public sector - allowing for greater transparency.

Question 3 & question 5.

How should decisions involving AI be reviewed and scrutinised in both public and private sectors? Are current options for challenging the use of AI adequate and, if not, how can they be improved?

To what extent is the legal framework for the use of AI, especially in making decisions, fit for purpose? Is more legislation or better guidance required?

- **The ability to challenge automated decisions involving AI is key to building trust in the technology.** In many sectors, there are already standards of care and processes to contest decisions regardless of how they were reached, but where AI is used in certain contexts, it is important that people can understand how and why decisions were made, with clear routes to challenge them.
- The primary legal route that enables people to do so is Article 22 of UK GDPR, which provides individuals with the right not to be subject to a decision based solely on automated processing which produces "legal or similarly significant" effects. As regulators set out AI-specific rules in their sectors, it is vital that they do so in a coordinated manner to avoid confusion or conflict between them.
- We were pleased to see the previous UK Government's draft Data Protection and Digital Information Bill aims to modify – rather than remove entirely – Article 22. Recasting it as a right to specific safeguards has the potential to make it a less ambiguous and more effective and accessible right.
- Beyond Article 22, the Government has proposed making 'clarifying routes to redress or contestability' a cross-sectoral principle for AI regulation. This is a sensible decision – however, it will be important for regulators to align with the Government's steer that rules must be proportionate, context-specific and designed to capture instances where decisions have produced a material impact on people's lives or rights – in line with the approach taken under Article 22.
- **It will also be important to balance the trade-offs between these principles.** On occasion, the distinction between, and applicability of, different governance regimes will require clarification. For instance, the principle of *'fairness'* has legal grounding in the UK GDPR where personal data is concerned, however in other frameworks fairness may have lesser standing as guidance, as an operating principle, or may not exist at all.
- The Government currently plans to put these cross-cutting principles on a voluntary footing, which makes sense today, but we believe these principles and their legal standing should be kept under review, and updated as necessary.

Question 4.

How should the use of AI be regulated, and which body or bodies should provide regulatory oversight?

- **We believe that it is most effective to regulate the use of AI, rather than the technology itself.** We are very supportive of the UK Government's plans to do so via a context-driven approach - with existing regulators providing oversight in their domains, rather than mandating horizontal and sector / application-agnostic requirements like the EU's AI Act now proposes.
- AI regulation needs the following key components:
    - **Deep domain expertise of the contexts in which AI will be used**. As an increasingly general purpose tool, AI will be used in many different ways across the entire economy - in healthcare, financial services, transport, energy and beyond. The challenges and opportunities it presents will be shaped by the specific use case and sector in which it's deployed. Existing regulators are therefore best placed to lead the design of tailored regulatory responses, based on their deep domain expertise.
    - **A clear focus on applications.** The risk of actual harm occurring will most commonly manifest via the deployment of AI systems and where they materially impact the rights and interests of people. Regulation should therefore initially focus on applications, rather than AI research - as per the EU's approach. This will also ensure that the UK's framework continues to facilitate the use of AI in research - including for scientific discovery - which is often curiosity-driven and exploratory in its early stages.
    - **An understanding of supply chain complexity**. AI supply chains are increasingly complex. New applications often build upon many different AI systems, including APIs, underlying base models, and open-sourced libraries and datasets, sometimes beyond the awareness or control of original developers. The number and diversity of actors within AI supply chains can also differ greatly between different sectors and applications. Responsibility for any requirements should therefore be tailored based on the specific context of use, and distributed across the supply chain based on which aspect is most likely to lead to harm *and* an actor's practical ability to comply given the structure of the market and the nature of their contribution.
    - **A proportionate framework for weighing benefits and harms**. Risk-based approaches to regulation have clear logic: one size rarely fits all. Routine, low-risk applications should be held to different standards to those that are high-risk. For example, the risk of harm from facial recognition is very different when it is used for public surveillance compared with unlocking a phone. **An often overlooked point is that risk-based regulatory approaches should be equally informed**

by *the risks of not using AI.* When evaluating AI-related risks in their domain, regulators should compare them against existing systems and the status quo. There are many limitations to existing tools, products and services that do not use AI. In some circumstances the risks posed by the status quo may be greater than those involved in using an AI system, either to the population overall or to specific groups.

- **A grounding in existing work and best practice**. A range of regulators are beginning to develop responses to the use of AI within their sector. Notable examples include the ICO's efforts to develop guidance on approaching challenges at the intersection of AI and data protection; and the Medicines and Healthcare products Regulatory Agency's exploratory work on the future of medical devices, including those that use AI. A sector-based approach will help to encourage and empower - rather than constrain - the evolution of these promising initiatives.

- **Despite these important strengths, there are challenges associated with this context-driven approach, delivered by existing regulators, which will need to be carefully managed.**
- A context-driven approach increases demands on the capacity, expertise and cooperation needs of regulators. There's also a risk that it may lead to overlapping mandates or a high degree of variance in how AI is regulated across sectors, adding complexity to business - particularly those operating across multiple sectors.
- For it to be effective, we believe government should:
  1. **Ensure sufficient investment in existing regulators' capacity,** with a focus on two areas:
     - **i) In-house expertise and understanding of AI.** The goal should be for them to sufficiently understand AI's primary capabilities and limitations, how and where it is being deployed in their sector or domain, and capacity to monitor new risks or opportunities arising from AI's use on an ongoing basis. Agencies like the Information Commissioner's Office (ICO), the Competition and Markets Authority and the Financial Conduct Authority have led the way in starting to build these capabilities - enabling them to more effectively understand how to balance the risks and opportunities of AI and other emerging technologies - and provide a successful model for other regulators to follow.
     - **ii) Innovative regulatory approaches.** Regulators should commit to experiment with and develop innovative approaches to regulation in their area. We welcome the government's emphasis on adaptability and the steer to regulators, where appropriate, to consider guidance and voluntary measures in the first instance. The UK already has helpful expertise and resources to draw on, such as the Regulators' Pioneer Fund,

the Regulatory Horizons Council (RHC), and the experience of agencies such as the FCA and the ICO with running regulatory sandboxes.

2. **Explore options for a new 'AI Governance Centre'** to provide guidance, resources and coordination capacity for regulators - whilst preserving their independence and flexibility to develop their own tailored approaches. Its key functions could include:
    - Convening, facilitating and incentivising regulatory collaborations around key AI issues - building on the great work of the Digital Regulation Cooperation Forum (DRCF).
    - Conducting cross-sector risk-mapping, regulatory gap analyses and horizon scanning.
    - Housing the new AI Standards Hub's work to coordinate the UK's input to global technical standards for AI.
    - Driving efforts to deepen and broaden the social dialogue on AI governance - and fostering greater participation.
    - Leading international collaboration, as well as seeking to learn from others, and benchmarking the UK's own framework against them.

- A new AI Governance Centre should build on and draw from the specialist capabilities and expertise already present in existing organisations - the UK's Centre for Data Ethics and Innovation (CDEI), the Digital Regulation Cooperation Forum, the Alan Turing Institute. All are well-placed to perform many of the functions that we think the Centre should deliver - and in many cases are already doing so in early form. The AI Standards Hub's work could also be housed within the new Centre as a key workstream, with the British Standards Institution an essential partner.
- Careful thought will need to be given to which functions are best centralised and which are best left to individual regulators - it's vital, for example, that data science expertise is not pooled centrally at the expense of building the capabilities directly in key regulators (see our point 1 above). Any new institution should be genuinely additive, target gaps in current capacity and also serve to simplify what is already a complex governance ecosystem.

Question 6.

What lessons, if any, can the UK learn from other countries on AI governance?

- **AI governance and regulatory efforts are accelerating globally.** It is important that the UK seeks to learn from and build on best practice from around the world, whilst also striving to encourage interoperability to the extent possible. Moreover, the UK's strength in AI gives it an influential seat at the table. There is a clear opportunity for the UK to act as a global convener, and lead some of these important conversations. For example, the UK could develop new AI-related standards in domains where the UK has particular strength or ambitions, such as health and life sciences, climate tech and fintech - then promote these through international fora.
- One place the UK should look for inspiration is the approach being taken by the US's National Institute of Standards and Technology to develop a voluntary **AI Risk Management Framework.** A new AI Governance Centre should seek to develop an equivalent AI Risk Management Framework for the UK - and seek to replicate the inclusive, iterative and open approach NIST has taken to its development. This could serve as a shared resource and common framework for a range of actors in the AI ecosystem to manage AI-related risks - as defined by the proposed cross-sector principles.
- The UK should also establish a close bilateral dialogue with the US on this topic - learning from and shaping the evolving framework in the US. The goal should be to establish mutual equivalence in approaches. The UK and US could also seek to lead a discussion, hosted by the OECD, to encourage broader OECD country collaboration on risk management standards - focusing on user needs and securing input from underrepresented groups in society. This would help to pool global resources on AI governance capabilities - including standards, benchmarks and analysis tools - and reduce potential barriers to trade that could otherwise emerge.

(November 2022)