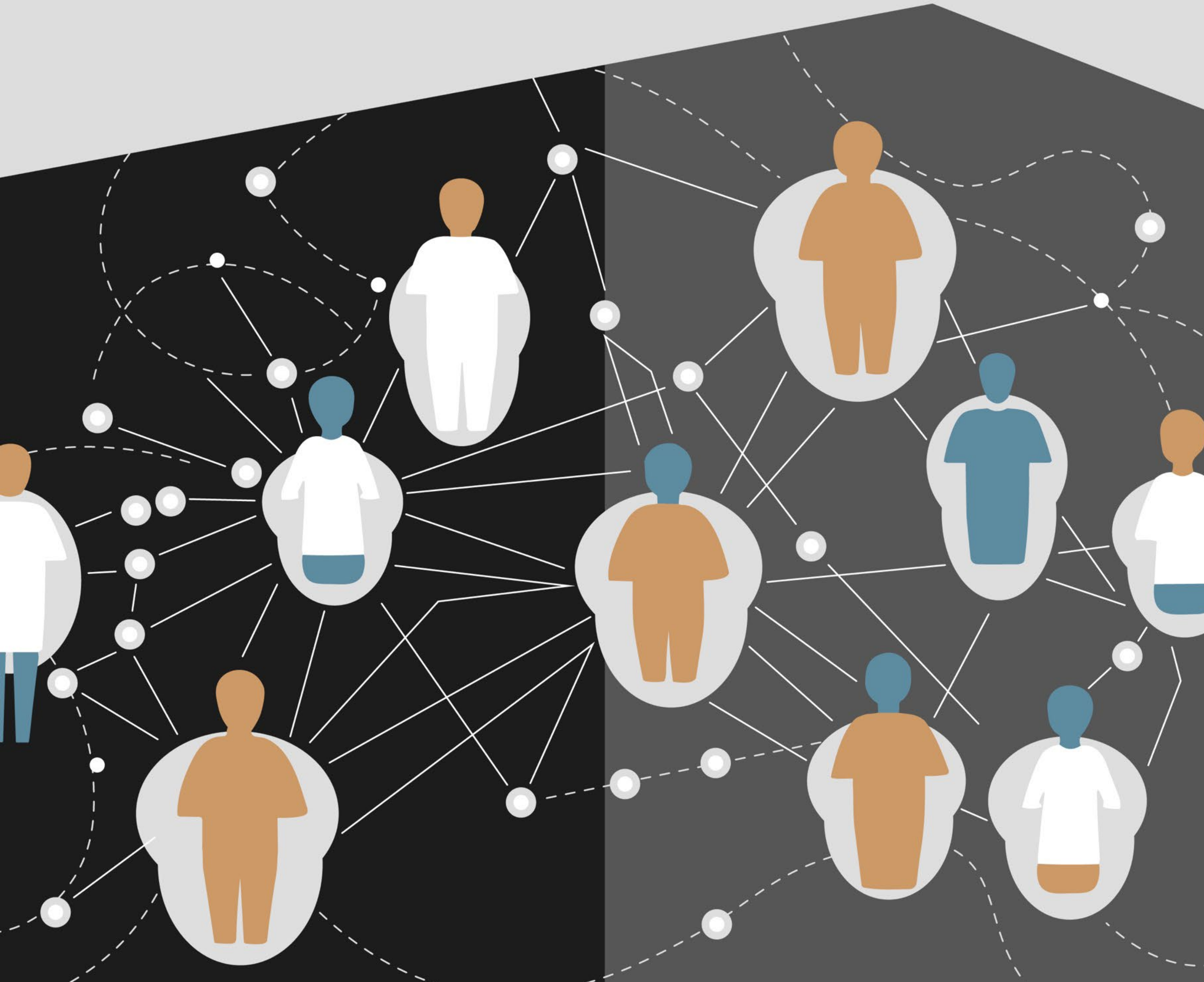
The Alan Turing Institute

September 2023

The Alan Turing Institute's response to the Large Language Models Inquiry: Call for Evidence



This document sets out The Alan Turing Institute's response to the House of Lords Communications and Digital Committee's Large Language Models Inquiry: Call for Evidence. The response synthesises the perspectives of researchers at the Turing with expertise and interest in the area of Large Language Models.

Table of Contents

Introduction and Summary 3

Responses to the questions asked in the Call for Evidence 7

Section 1: Capabilities and trends 7

1. How will large language models develop over the next three years?.. 7

1a. Given the inherent uncertainty of forecasts in this area, what can be done to improve understanding of and confidence in future trajectories? 9

2.What are the greatest opportunities and risks over the next three years? 10

2a. How should we think about risk in this context? 14

Section 2: Domestic regulation 16

3. How adequately does the AI White Paper (alongside other Government policy) deal with large language models? Is a tailored regulatory approach needed?.. 16

3a. What are the implications of open-source models proliferating?... 18

4. Do the UK's regulators have sufficient expertise and resources to respond to large language models? If not, what should be done to address this?... 20

5. What are the non-regulatory and regulatory options to address risks and capitalise on opportunities?... 20

5a. How would such options work in practice and what are the barriers to implementing them? 24

5b. At what stage of the AI life cycle will interventions be most effective? 25

5c. How can the risk of unintended consequences be addressed?... 27

Section 3: International context 27

6. How does the UK's approach compare with that of other jurisdictions, notably the EU, US and China? 27

6a. To what extent does wider strategic international competition affect the way large language models should be regulated? 29

6b. What is the likelihood of regulatory divergence? What would be its consequences? 30

References 32

Authors 35

Introduction and Summary

Since the start of this century, increases in processing power, in particular the use of Graphics Processing Units (GPUs), and the widespread availability of large and curated datasets have driven important advances in AI and machine learning, particularly the sub-field of deep learning. Foundation models are the latest example of these factors leading to powerful new capabilities that can be adapted to various purposes (hence 'foundation'). Large Language Models (LLMs) are a subset of foundation models focused on language.[1] LLMs are often described as a form of generative AI, i.e., foundation models that create new content, such as text, images, audio or video.

LLMs have been a subject of interest to the AI research community for years prior to ChatGPT's launch in November 2022. However, ChatGPT marked the first widely available release of an intuitive general purpose tool based on a LLM, and thus precipitated an explosion of interest in LLMs from the public, media, policymakers and industry.

The Turing welcomes this inquiry as a chance to focus policymakers' and parliamentarians' attention on the immediate opportunities and risks posed by LLMs, and the urgent need to implement policy to manage identified risks without sacrificing the opportunities that LLMs offer across many sectors of the economy. While this submission focuses specifically on LLMs, many of the same considerations discussed in relation to LLMs also extend to other generative AI models as they share many of the same challenges and opportunities.

Capabilities and trends (over the next three years)

1. Predicting future technological breakthroughs is challenging for numerous reasons including a lack of transparency in research and development practices. However, the application of *existing* LLM technology to new use cases could itself produce major impacts over the next few years.
2. This includes in a research context, where LLMs, if employed effectively and responsibly, offer numerous potential benefits for research and innovation. The Alan Turing Institute's strategy aims to take advantage of this across critical areas including health, environment and sustainability, defence and security, and digital society and policy.
3. The immediate risks posed by LLMs are well documented in the academic literature. These risks may stem from an uplift in capability for malicious

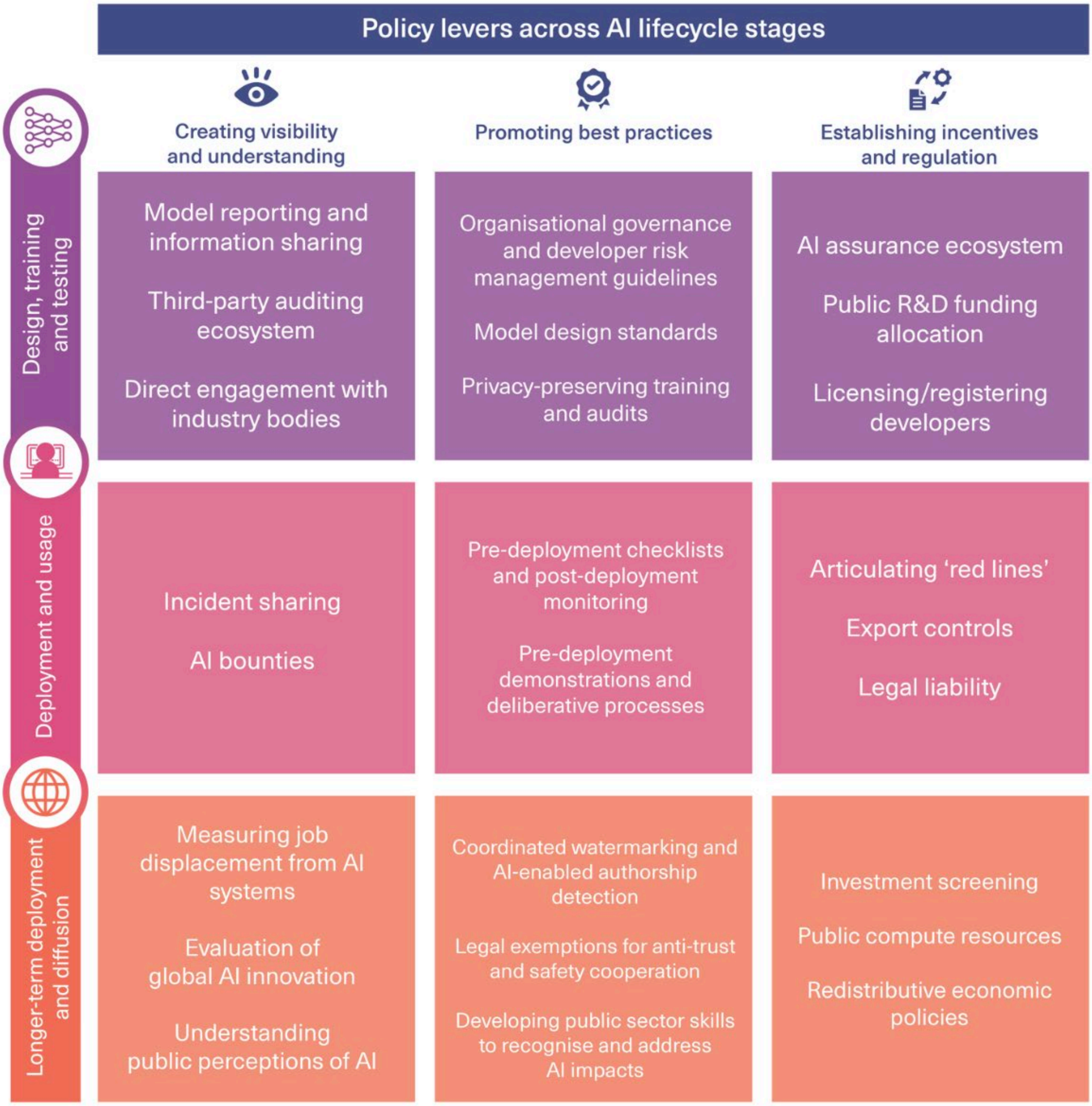
[1] Here we occasionally refer to foundation models, 'generative AI', and LLMs concurrently because they share many of the same risks and governance challenges.

actors, giving them new routes through which to undermine the UK's digital, physical and political security. Some examples include malicious actors integrating LLMs into fraud and cybercrime activities, using LLMs to generate information about building weaponry and developing attack plans, and significantly altering the speed and scale of mis/disinformation operations intended to undermine liberal democracy. Risks may also stem from irresponsible deployment, resulting from LLMs' presentation of factually incorrect information as true ('hallucinations'), leaking of private information, and experimentation with LLMs by individuals and organisations conducted without malicious intent, yet without adherence to clear guidance and safeguards, leading to significant unintended consequences.

4. There are also wider systemic and societal risks beginning to emerge, whose impact may only be fully realised in the longer term. These include perpetuating or amplifying existing social biases and discrimination, causing environmental harm, and adversely impacting workers across the AI lifecycle.

Domestic regulation

5. To address the diversity of risks posed by LLMs, and generative AI more broadly, we suggest the government enhances its sector-based approach to AI regulation (proposed in the AI White Paper) by clarifying its central function; tailoring proposed regulatory sandboxing initiatives to LLMs; and prioritising placing the cross-sector principles on a statutory footing.
6. There are also legitimate decisions to be considered about the funding and capacity needed in existing regulators to address novel risks from LLMs and other foundation models, particularly among smaller regulators.
7. A suite of regulatory and non-regulatory options beyond those proposed in the White Paper may be useful in mitigating risks from LLMs. The former would be bolstered by mandating auditing processes and methodologies which focus on bias mitigation and explainable design, and developing LLM-specific standards. The latter would benefit from interventions focused on model reporting and information sharing, systematic incident sharing and analysis, pre-deployment checklists, demonstrations and deliberative processes to improve public understanding of LLMs, and post-deployment monitoring (particularly in domains involving a higher risk of accidents or misuse).
8. Proposed interventions should take place throughout the AI lifecycle, although for existing LLMs an emphasis should be placed on post-deployment interventions to mitigate known risks and harms.
9. The following table summarises the main policy levers available to mitigate AI risks across the AI lifecycle, extracted from the recent Turing paper on 'Strengthening Resilience to AI Risk', published by the Turing's Centre for Emerging Technology and Security in partnership with the Centre for Long-Term Resilience (Janjeva et al., 2023):



Policy levers across AI lifecycle stages

	Creating visibility and understanding	Promoting best practices	Establishing incentives and regulation
Design, training and testing	Model reporting and information sharing Third-party auditing ecosystem Direct engagement with industry bodies	Organisational governance and developer risk management guidelines Model design standards Privacy-preserving training and audits	AI assurance ecosystem Public R&D funding allocation Licensing/registering developers
Deployment and usage	Incident sharing AI bounties	Pre-deployment checklists and post-deployment monitoring Pre-deployment demonstrations and deliberative processes	Articulating 'red lines' Export controls Legal liability
Longer-term deployment and diffusion	Measuring job displacement from AI systems Evaluation of global AI innovation Understanding public perceptions of AI	Coordinated watermarking and AI-enabled authorship detection Legal exemptions for anti-trust and safety cooperation Developing public sector skills to recognise and address AI impacts	Investment screening Public compute resources Redistributive economic policies

International context

10. The UK's proposed sector-specific approach to AI regulation is a "middle of the road" approach between the China and the EU's stringent regulations and the US's lighter-touch approach. It is important to recognise the mutually reinforcing relationship between domestic and global policy interventions: by being proactive with domestic AI policy implementation, the UK will be better placed to advocate for those policies globally, which will in turn generate further credibility and support for the UK's domestic AI ecosystem.
11. Given the multinational scope of LLM developers, it is important to address anticompetitive measures and reduce compliance costs to ensure that innovation can continue to flourish among small and medium sized entities.
12. The Government should consider the extent of possible regulatory divergence between the UK and EU (and where relevant, other jurisdictions), which could increase cross-border compliance costs and risk stifling innovation.
13. The UK Government's forthcoming AI Safety Summit presents an important opportunity to focus global collaboration on these challenges, particularly given the international nature of the development and use of LLMs.

The Alan Turing Institute is collaborating widely across academia, government, civil society and industry to maximise the potential benefits of AI development while driving research efforts to better understand and mitigate risks. The Institute welcomes this Inquiry and will continue to engage in open and inclusive dialogue on these critical policy issues.

Responses to the questions asked in the Call for Evidence

Section 1: Capabilities and trends

1. How will large language models develop over the next three years?

Predicting future technological breakthroughs is challenging. However, major leaps in technological capability are not needed to see significant impacts through applications of LLMs to new uses. Although we do not anticipate a major technological breakthrough (except from integrating more modalities or other related technologies into the same models), predicting such breakthroughs is notoriously difficult. This is because of the lack of information available about AI development by the companies responsible for the most powerful models. However, a major technological breakthrough is not needed to see significant impacts, including increased efficiency, scale of deployment, sophistication, and integration, which can unlock new capabilities using existing core LLM technology and other generative AI technology.

Regardless of technological breakthroughs, there could be significant developments for LLMs as we do not understand how scaling works in practice. It is difficult to predict the behaviour of models larger than what we have now, and each size increase brings new unknowns. A lot of the novel capabilities we have seen in the last year have come from doing just 'more' of the same, i.e., training larger models with more data but of the same type and in the same way. The emergence of these capabilities is not trivial: it is not clear why a large model can write, for example, convincing poetry on a given topic, but a smaller one cannot. More critically, we cannot predict what even larger models could do.

The proliferation of open-source models is leading to a growing uptake by developers and researchers. Open-source models have begun to proliferate and will continue to do so. Important inflection points in this trajectory have been the release of GPT-2 by OpenAI in 2019 and the development of the Transformers library and Hugging Face Model Hub by Hugging Face. Since then, numerous open-source LLMs, libraries and tools have been released, leading to a dynamic environment as researchers and organisations build upon and improve existing models and contribute to the development of new ones. These are likely to spur innovation in the private sector and research across sectors, though it is important to note both the benefits, risks and means to manage these.

Integration with other systems will increase. While generative AI systems such as ChatGPT (text) and StableDiffusion (images) are already widely used on their own, there is a considerable effort to extend them and combine them with other systems. So far, most of these efforts have been prototypes and proof-of-concepts. For example, researchers have worked to integrate LLMs into control robots and generate assets in popular 3D modelling software. Most notably, OpenAI provides add-ons to ChatGPT which allow it to browse the internet, run code, perform real world actions such as placing orders on the internet and schedule meetings with real people. In the next few years these integrations will likely grow in sophistication.

Developers will focus on reliability and truthfulness of information. There are efforts towards developing more reliable and trustworthy models (e.g., models that provide sources when offering factual information), which can increase their practical utility. As one of the primary limitations of LLMs is their propensity to 'hallucinate', described in further detail under question 2, progress towards safety and reliability of LLMs, and other generative AI systems, will be critical. The Turing welcomes the AI Safety Summit as an important means to focus global collaboration on these challenges, particularly given the international nature of the development and use of LLMs.

Development may be limited by hardware or computing power constraints, though models will likely continue to grow in scale. In the next few years, hardware will likely become even more valuable and more difficult to access—for example, A100 chips, which are commonly used in AI applications, are already providing difficult to obtain in Europe, and it is possible that advanced GPUs could be designated as dual use by the US under the Chips Act, making access to them more challenging. Moreover, despite the difficulties accessing compute and contrary to what some companies have publicly announced, we expect that the scaling towards larger models will continue.

Increase in LLM-generated data may limit further development. The amount of AI-generated data on the internet is increasing and will only accelerate, which means that the fraction of AI-generated data used for training future versions of LLMs will also increase. Ingesting AI-generated data can negatively impact the quality of AI systems (Alemohammad et al, 2023). Therefore, a decreased access to high quality data may be a limiting factor for the further development of the field.

Risk assessment could remain challenging without further research into evaluation tests. Determining the risks posed by more capable AI is, with the tools currently available, very challenging. Without further development of tools such as evaluation tests it will be difficult to assess the capability of new AI models. Providing academia with early access to frontier models could enable a wider audience of experts to input into discussions and evaluations of new LLMs.

1a. Given the inherent uncertainty of forecasts in this area, what can be done to improve understanding of and confidence in future trajectories?

Government could establish a voluntary information sharing regime with frontier AI labs. The UK government shared in June 2023 that it secured pledges from three leading AI labs to grant 'early access' to their models. This pledge is to be welcomed and could be operationalised through the establishment of an information sharing regime, connecting labs developing foundation models (see full details of a proposed pilot model by Mulani et al., 2023). Such a regime between AI system developers and government bodies may be a useful method to provide the UK government foresight into emerging risks and opportunities. Relevant categories of information could include models' intended functionality, levels of compute usage during training, evaluation against performance benchmarks, and information on training datasets. This could be provided to a central body in government, for dissemination to relevant policymakers and regulators.

A key regulatory challenge is that many of the developers building the most powerful models are based outside of the UK. The UK government could demonstrate an effective and systematic means of information sharing that can be replicated in other countries, in concert with prioritising international collaboration, discussed below.

Sharing data with academics could help government understand future trajectories. Currently, most public (non-industrial) specialised AI expertise resides in academia, rather than in government, and thus industry data should be shared more widely so academic researchers can analyse it to understand future trends. Given this data is commercially sensitive, the government would need to take the appropriate non-disclosure and conflict of interests considerations when facilitating data sharing with academics.

Convene international collaboration, premised on an understanding of shared risks. LLMs will continue to develop in an international context, with many of the developers responsible for the most powerful models based outside of the UK. However, the UK is not alone in wanting to mitigate the risks of these technologies. The UK also has much to bring to the table, from its leading researchers and ethical frameworks as well as its well-respected regulatory regime. The UK should therefore prioritise the cooperation and coordination of key partners, including the US, Canada and the EU to develop common approaches to the mitigation of risks shared across borders, including common approaches to data gathering. The upcoming AI Safety Summit is a good opportunity to kickstart this work.

Government could collect and analyse proxy data on LLM developments, in particular purchases of GPUs. Even with the tendency towards industrial secrecy, there is a significant quantity of data accessible to government which can be analysed to predict and increase confidence in the trajectory of future LLM developments. This includes non-public data which is available to the government such as large purchases of GPUs. Although this information is not systematised and is generally difficult to analyse, better aggregation, tracking and analysis can help improve understanding of and confidence in future trajectories.

2. What are the greatest opportunities and risks over the next three years?

<u>Opportunities</u>

Accelerating academic and industrial research. LLMs inherently do not possess unique knowledge that is not already accessible on the internet. However, they appear to have an impressive capacity to amass, distil, and connect knowledge from a variety of sources. LLMs can identify links and correlations within vast amounts of data that would otherwise go unnoticed, facilitating a more holistic understanding of complex systems, which has the potential to drive innovation and growth. This capacity for consolidation can significantly speed up both academic and industrial research.

At the Turing, we are particularly interested in working with domain partners to explore the responsible application of LLMs to key science and innovation areas in our strategy, including environment and sustainability, health, defence and security, and digital society and policy.

Enhancing accessibility of information in a variety of contexts. LLMs provide an innovative approach to information retrieval. Users can provide a description or explanation of their query, and the model can iteratively refine its understanding of the information being sought based on the user's feedback. In the near term, LLMs could form the basis of knowledge management tools that enhance productivity in a variety of professional sectors, from consultants to nurses, enabling workers to access relevant information faster. Of course, risks caused by ‘hallucinations’, already mentioned but discussed in further detail shortly, need to be considered.

Enhancing productivity for software tools. Soon, many popular software systems could have natural language interfaces which run alongside the traditional point-and-click interface. This can remove or improve the learning curve for advanced software systems used in design, engineering, finance, enterprise management and others. Currently, such systems are rudimentary and often unreliable, and require constant supervision by a skilled practitioner. In the next few years, however, they will likely become more reliable and require less supervision. Such levels of integration and functionality may bring economic benefits and productivity improvements.

Increased productivity of writing and procedural tasks across sectors. One of the most immediate uses of LLMs is as a ‘productivity assistant', which can automatically complete sentences, proofread emails and documents, and automate certain repetitive tasks. As sophisticated text processors, LLMs can undertake tasks such as reformatting comprehensive documents and maintaining consistency in style. This could potentially reduce the time spent on manual editing, although the results would still require human review for accuracy.

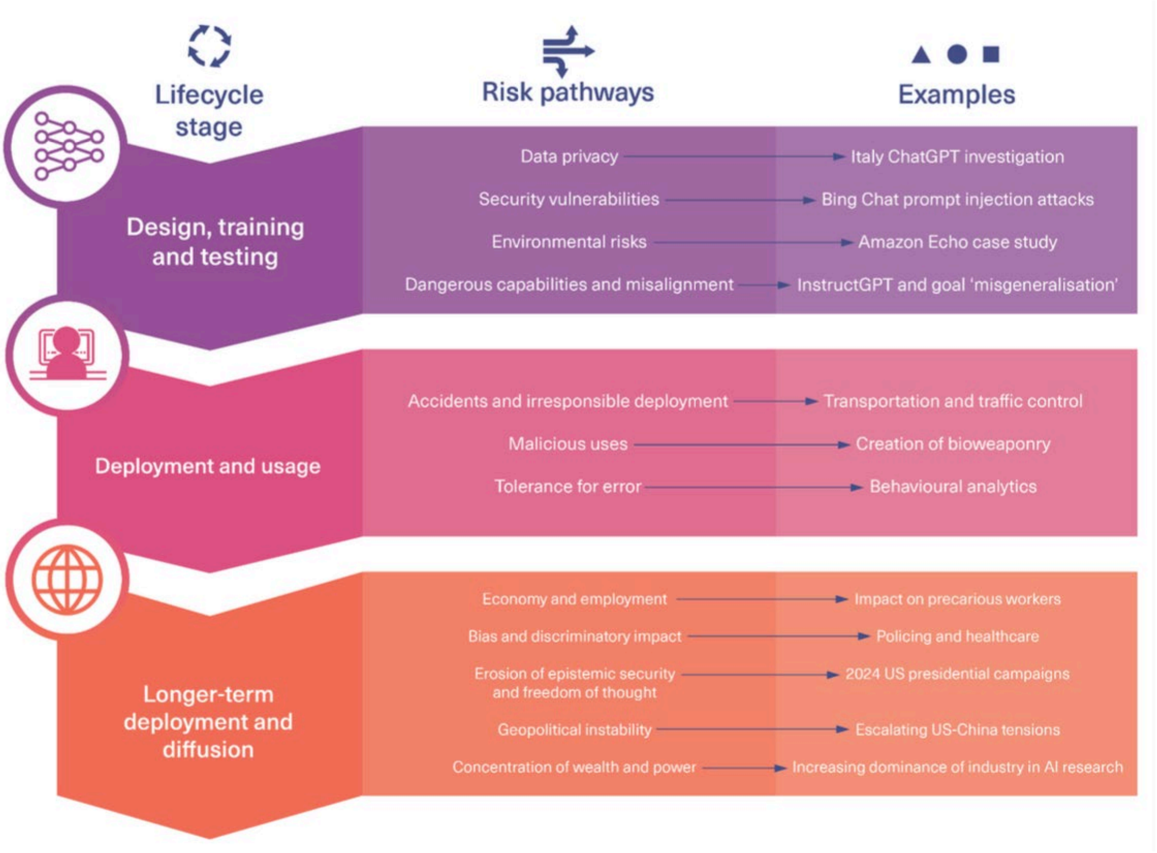
Software development. In the context of software development, there has been significant interest in experimentation with the use of LLMs to assist with writing code. However, as of now, the outputs of code assistant LLMs are sometimes inaccurate, suboptimal or not secure (Pearce et al., 2022), often necessitating human oversight and correction. While we would expect this to improve over time, this currently limits productivity gains as AI-generated code can be more challenging to debug.

Personalisation of services, from entertainment and marketing to education. LLMs' capacity for rapid creation of customized content allows for new technologies which can personalise experiences, services, and products. In entertainment, for instance, LLMs and other generative AI could facilitate the generation of videos and music on demand, based on specific user preferences. This is leading to questions about the nature of human art as well as raising copyright concerns, discussed under 'risks'. In marketing, we can expect to see increasingly personalised advertisements. In education, LLMs could deliver a tailored experience by synthesising learning materials for the exact needs of the individual student and adapting to students with neurodivergence or learning disabilities. The Turing's submission to the Department for Education's Generative AI call for evidence discusses LLMs in the context of education in further detail.

Risks

The future of LLMs contains numerous similar 'unknown unknowns' or unanticipated challenges. These may arise from the complex interaction of these technologies with social, cultural, and economic dynamics.

Risks from LLMs could arise at any stage in the AI lifecycle: from design, training and testing; immediate deployment and usage; and longer-term deployment and diffusion. The following mapping is not intended to be exhaustive, but rather to give an overview of the range of potential risks that can arise:



Lifecycle stage	Risk pathways	Examples
Design, training and testing	Data privacy	Italy ChatGPT investigation
	Security vulnerabilities	Bing Chat prompt injection attacks
	Environmental risks	Amazon Echo case study
	Dangerous capabilities and misalignment	InstructGPT and goal 'misgeneralisation'
Deployment and usage	Accidents and irresponsible deployment	Transportation and traffic control
	Malicious uses	Creation of bioweaponry
	Tolerance for error	Behavioural analytics
Longer-term deployment and diffusion	Economy and employment	Impact on precarious workers
	Bias and discriminatory impact	Policing and healthcare
	Erosion of epistemic security and freedom of thought	2024 US presidential campaigns
	Geopolitical instability	Escalating US-China tensions
	Concentration of wealth and power	Increasing dominance of industry in AI research

For further discussion, see 'Strengthening Resilience to AI Risk', published by the Turing's Centre for Emerging Technology and Security in partnership with the Centre for Long-Term Resilience (Janjeva et al., 2023).

Accuracy and performance

Harms caused by 'hallucinations'. A significant risk associated with generative models is their propensity for 'hallucinations', or the creation of plausible-sounding but inaccurate or fabricated information. If an output is relied upon without fact-checking, it can lead to significant errors in decision-making, with resulting damage to individuals' or organisations' reputations. At the same time, if outputs are constantly being fact-checked, the utility of the tool and associated productivity gains will be more limited.

Data contamination. A compounding concern is the potential degradation of output quality as AI systems consume data generated by other AI systems. This suggests that as we lean more heavily on LLMs, the performance of these models may paradoxically decline.

Enabling malicious actors

Enabling tailored mis/disinformation on a large scale. LLMs could transform the speed and scale at which malicious actors generate mis/disinformation, potentially flooding the digital public square with misleading and non-factual content. This mis/disinformation can be tailored, as LLMs can collate disconnected personal data scattered across the internet and construct detailed profiles of individuals with speed and at a scale currently close to impossible. This could divert attention away from key issues, encourage the persistence of echo chambers with poor epistemic norms and allow malicious actors to fake the hallmarks of trustworthy information sources. A worst-case scenario would leave democratic societies like the UK unable to sustain informed electorates, while giving authoritarian regimes greater tools of control and suppression.

Enabling impersonations, scams and cyberattacks. LLMs are driving three key improvements which are changing the fraud and cybercrime landscape: speed and efficiency of creating a scam from scratch through to exploiting victims; convincingness; and a reduction in the technical competence required to do so. The ability of LLMs to respond to messages in context and adopt specific writing styles are crucial to enhancing the quality of scams, while developments in the field of autonomous agents may lead to a step change in quantity.

Guiding the creation of various attack mechanisms. Sophisticated LLMs may have the potential to guide the creation of biological, nuclear and conventional weapons, or launch cyberattacks, dramatically increasing their efficiency and reach while lowering the required sophistication of the malicious actor. For example, in just one hour of use, chatbots 'suggested four potential pandemic pathogens, explained how they can be generated from synthetic DNA […] supplied the names of DNA synthetic companies unlikely to screen orders, identified detailed protocols and how to troubleshoot them, and recommended that anyone lacking the skills to perform reverse genetics engage a core facility or contract research organisation' (Soice et al, 2023).

Copyright

Significant copyright concerns. Generative forms of LLM models often inadvertently reproduce copyrighted content. This capability can lead to the creation of outputs that infringe on existing copyrights, which has already instigated several high-profile legal disputes. The emerging tension between AI's replication capabilities and intellectual property rights is likely to escalate.

Social, economic and environmental impacts

Perpetuate and amplify social bias. One particularly pressing concern is the potential for these systems to perpetuate and amplify existing societal biases. Given that LLMs are generally trained on immense amounts of real-world data, they can inadvertently learn and reproduce patterns of discrimination present in their training data. Moreover, the individual decision-making process of AI systems can be challenging to scrutinize, rendering bias detection and rectification difficult.

Environmental sustainability. The initial training of LLMs has significant environmental impacts due to the huge energy consumption this entails (Luccioni et al, 2022). The larger a model is, the greater the energy consumption required to train it. Thus, the interest by developers in building larger and larger models suggests a trajectory of increased energy consumption, and linked emissions. Additionally, as with the wider computing industry, biodiversity loss, chemical waste, and water use are significant issues related to the extraction of raw materials for AI systems and for cooling data centres at the beginning of the AI supply chain.

However, it is worth noting that the large upfront energy costs to train a model can, due to their great potential for applicability in numerous contexts, be amortised across their many uses. These uses, involving the direction of existing models to specific applications, themselves do not necessarily incur large energy use. The research community is also actively working on ways to reduce the carbon footprint of LLMs. There are several promising approaches, such as 'load shifting' to move intensive computer processes to regions or times of the day to align with renewable energy supply. As these approaches are developed and implemented, the carbon footprint of LLMs is expected to decrease.

Exploitative practices. Exploitative labour can play a part in generative AI supply chains—from illegal scraping of data to train models, to data labelling, content flagging and other activities taking place in different countries with lower levels of labour protection (see).

Market consolidation leading to concentration of economic power and social influence. Growing reliance on the capabilities of commercial LLM technologies owned by a small number of companies could precipitate the consolidation of economic power due to their control of data, compute, and model engineering infrastructures.

2a. How should we think about risk in this context?

As illustrated previously, risks from LLMs could emerge at all stages in the AI lifecycle. Moreover, it may be impossible to predict the full spectrum of risks that could arise from the deployment of LLMs in different sectors. For this reason, policy interventions must build resilience to risks throughout each stage of the AI lifecycle, to mitigate known harms from AI, and anticipate and prevent future risks.

Discussions about risk in an LLM context should distinguish between immediate and speculative concerns. Driven by the interest and popularity of currently widely available LLM platforms, there is often a conflation of LLMs with AI more generally. Linked to this, there has also been a blurring of attendant risks, which can broadly be divided into two categories: those that affect us now, and speculative risks and existential concerns. Coverage and commentary often focus on this latter category. Although worthy of consideration, this should not distract from immediate concerns including perpetuating bias, misinformation and enabling malicious actors.

Unpredictability and speed. The adaptable nature of these models can result in outcomes that were not anticipated nor intended by the developers. As the technology continues to evolve, LLM-generated harms are also likely to propagate at faster speeds and in greater quantities. This suggests that additional governance measures focused on earlier stages of the AI lifecycle – to manage the way that certain AI models are developed and initially deployed – will be needed to mitigate the full range of potential harms.

Asymmetrical standards for development and deployment between legal and illicit entities. Legitimate applications tend to be subjected to rigorous quality and performance checks before release, making their product development cycles longer and fraught with uncertainties. In contrast, illegitimate entities exploiting AI for nefarious purposes do not need high performance and it does not matter if they are inaccurate more often. Realising the opportunities of legitimate AI applications might take longer and be more uncertain than realising the risks of malicious use.

Risk is not distributed equally across society. Although the overall impact of AI across society is likely to be beneficial, this effect will be unevenly distributed across sectors and demographics, risking the amplification of inequalities.

Adaptable, domain specific approaches to risk management are important. Moreover, the general-purpose nature of LLMs, along with generative AI more widely, implies its potential impacts span across sectors, necessitating a hybrid approach to risk management. Centralised strategies rooted in a deep understanding of AI technology should be combined with domain-specific approaches to capitalise on sector-specific expertise. The dynamic nature of AI development, marked by rapid evolution and emergence of novel research problems, necessitates continuous adaptation in risk management strategies. These strategies should encompass ongoing data collection, regular reassessments, and iterative improvements, underlining the fluidity of the AI risk landscape.

Section 2: Domestic regulation

3. How adequately does the AI White Paper (alongside other Government policy) deal with large language models? Is a tailored regulatory approach needed?

As the Turing noted in our original response, the Government's AI White Paper builds on the strength of existing regulators, and creates a "principle-based, sectoral approach [which] is crucial to ensure that AI is regulated in the most effective and efficient way."

However, the fundamentally cross-sector nature of the risks posed by LLMs require close consideration in the context of this sector-based approach. The Government's proposed regulatory approach is primarily designed to regulate the outcomes of AI on a sector-by-sector basis; however, in the case of LLMs, and other foundation models, outcomes are diverse and span multiple sectors. In addition, assessing and regulating only high-risk applications of such technologies is impractical, since predicting the risk level of every possible foundation model use case is infeasible. As discussed previously, seemingly low-stakes applications may be used by bad actors to cause harm or may cause unintentional harm as a by-product of the model's design or development.

In particular, there is a possibility that risks posed by LLMs fall in the gaps between regulators or create uncertainty by falling within the remit of multiple regulators. Given the rapidly developing and cross-sectoral nature of LLMs, they are a prime example of a technology that may not be adequately covered by existing regulatory remits, necessitating cross-sectoral or horizontal coordination to address such gaps. Additionally, some risks may fall within multiple sectors and therefore different regulators may apply and interpret the White Paper's principles differently to these risks based on their existing statutory duties and remits, leading to conflicting monitoring and enforcement approaches. Regulators may produce contradictory guidance, resulting in uncertainty for regulated entities or encouraging them to search for paths of least resistance.

Recommendations to enhance the AI White Paper's regulatory approach

To address the risks LLMs pose, we echo our recommendations into the AI White Paper for the need to set out a clear plan for central coordination, the building of shared expertise and the prioritisation of putting principles on a statutory footing to ensure consistency.

These suggested improvements, while useful for general purpose AI technologies more widely, are set out below with particular reference to their application to the governance of LLM risks:

1. **Setting out a clear plan for the central coordinating function and ensuring it has access to relevant expertise, including on LLMs**
   Due to the cross-sectoral nature of LLMs, the need for central coordination in the UK's regulatory approach is an important priority. Although the central function outlined in Section 3.2.4 of the White Paper is tasked with such coordination, more information setting out how the central function will operate with regards to its powers, independence, and delivery mechanisms would be welcome. To ensure that the central function can play an effective role in addressing the risks posed by these models, the we suggest articulating a clearer plan for this function and how it will interact with UK regulators would be useful. The central function will need to be appropriately independent to avoid conflicts of interest when monitoring and assessing the effectiveness of the approach to regulating foundation models (and specifically LLMs). The central function will also require sufficient access to a range of expertise to effectively deliver on complex tasks such as risk assessment and will need to be sufficiently empowered to resolve conflicts and uncertainties between regulators.

2. **Tailoring regulatory sandboxes to support LLM applications**
   Regulatory sandboxing, which the White Paper's approach supports, will be an important step in the process of developing regulations that can work across-sectors, given the wide-ranging risks posed by LLMs, as well as in fostering regulatory cooperation. In this context, it will be important to focus on sandboxes that involve **multiple regulators across multiple sectors** to provide the adequate scope to address LLMs. A multi-regulator sandbox has been recommended in the recent Pro-innovation Regulation of Technologies Review, led by Sir Patrick Vallance, with this recommendation being noted by the Digital Regulation Cooperation Forum (DRCF) in their 2023/24 workplan.

3. **Prioritising placing the cross-sector principles on a statutory footing.**
   The White Paper outlines that cross-sectoral principles will be issued initially on a non-statutory basis, with no timeline announced for when a statutory duty may be introduced. Regulators will already be operating with different interpretations of the principles based on their existing mandates, and the principles will have to be implemented across a patchwork of varying regulatory powers. If each regulator is left to interpret the principles themselves, the developers or users of a foundation model such as an LLM may be faced with inconsistent or incomplete regulatory requirements. This could lead to a range of issues, including possibilities of arbitrage (e.g., firms opting for the regulatory path that offers the least resistance and stymying of innovation due to a myriad of conflicting rules and regulations). A statutory footing is recommended to provide regulators with the appropriate powers and mandates to enact tailored LLM regulations.

It is also worth nothing that the current iteration of the White Paper's approach (as well as other Government policy) does not fully address the following risks, which are present in the context of foundation models and LLMs.

The downstream nature of LLM uses and risks can create gaps in accountability and liability. Foundation models, including LLMs, are deployed across complex, non-linear supply chains, which creates a "many hands" problem for assigning accountability and a further problem for assigning liability to parties that can mitigate harms (Cobbe et al., 2023). Issues can arise from the adaptation of LLMs by different users that may lead to previously unforeseen risks that have not been anticipated by the model's creator. A lack of clear lines of accountability could result in situations where parties are not equipped to anticipate, identify, or deal with harms resulting from the use of an LLM. As outlined in our response to the White Paper consultation, clarification of liability and international coordination on these issues would be welcome.

Intentional misuse. The existing White Paper framework seems to focus on unintended harms rather than intentional misuse by malicious actors. In the case of LLMs, there is a substantial risk resulting from intentional misuse (as discussed in Question 2). Deciding where to allocate responsibility after a harm has occurred will likely not be sufficient in the case of extreme harms caused by groups that already anticipate imprisonment if they are caught. While the capabilities of these groups could drastically change with the emergence of generative AI more broadly, regulatory protection against these harms would be welcome.

Environmental sustainability. As described under the risks in response to Question 2, as LLMs become larger and more ubiquitous, the UK's regulatory approach can be used to encourage greater consideration of environmental sustainability, ensuring that we can balance the benefits and risks of these systems appropriately across the whole value chain.

3a. What are the implications of open-source models proliferating?

Open-source models are essential for democratising access to AI.

Open source models can be more easily tuned for specific purposes, diversifying opportunities for deployment, and to be adapted for languages other than English. They are therefore key for boosting innovation by small and medium-sized companies and startups. Similarly, they are a valuable tool to researchers in adapting models to applications across domain areas.

Open source models facilitate regulation and democratise access to AI. It is only through models that have been made accessible that we have public knowledge and understanding of how the models work and crucially, where they do not work well.

Most of the current research on LLMs and generative AI, including research on safety, security and alignment of these models, is based on openly available models and could not have been conducted without them.

The Turing is participating in the Open Source Initiative's ongoing work to collaboratively define "open source AI". One aspect of this is focusing on including the vast ecosystem of stakeholders in the development of AI: developers, dataset owners, the people represented in those datasets, and the – often under paid and unacknowledged – people in developing nations who moderate the content and outputs of state-of-the-art models as they are trained. Another centres on applying the SAFE-D principles derived from the national public sector guidance co-produced by Turing, "Understanding artificial intelligence ethics and safety" to open source AI.

We suggest the UK needs to be proactive in setting clear guidelines for governance that includes both open and closed source models.

Sharing elements of models as openly as possible is fundamental to our understanding of how these models work, and offer a greater ability to scrutinise the safety, security and alignment of these models compared to closed source systems, which are currently limited to private companies and organisations. While closed LLMs that are only accessible via an application programming interface (API), may offer the possibility of monitoring to detect attempts to cause harm using AI, there are already multiple examples of approaches undermining safeguards apparently built into closed source models. On the other hand, open source models are at greater risk of retraining to more harmful datasets with the intention to circumvent safeguards. We can expect bad actors to continue to exploit both avenues, in much the same way that hackers continue to attempt access to other computer systems despite regulations and cyber security efforts.

Closed and open source LLMs have different attributable risks, but neither is inherently a safer approach to LLM development. Guidelines on model governance should consist of an evaluation of potential harms of the model before any model is released. In particular the accountability pathway for decisions made by the humans creating the model should be transparently communicated for all models. Continual stakeholder engagement should also be a part of this process. Following the guidance of the people represented in the data and affected by the AI is fundamental in assessing who should make the decision about what is openly available and what should be protected for commercial competitive advantage.

4. Do the UK's regulators have sufficient expertise and resources to respond to large language models? If not, what should be done to address this?

Regulators should continue to build levels of expertise and readiness. Larger regulators (including members of the DRCF) have been building up significant expertise in the area of AI and digital technologies (for example, Ofcom in advance of the Online Safety Bill). This accumulation of expertise is necessary to understand the benefits and limitations of generative AI. Smaller regulators, however, may not possess the same levels of expertise. A pooled team of interdisciplinary experts could also assist regulators in addressing capability gaps. In addition to continuing to build expertise, appropriate levels of readiness are necessary to properly regulate technologies such as LLMs. As detailed in the Turing's Common Regulatory Capacity for AI Report, which was cited in the second version of the White Paper, regulatory readiness must be interrogated and scrutinised at three distinctive levels:

- **The individual level** (e.g., attitudes, perceptions, cognitive abilities, skills, and investments that enable individuals to embrace and integrate AI innovation and AI-prompted policy change);
- **The readiness of organisations** (e.g., the way that the institutional culture, the availability of resources, and the environment of policies, procedures, and collective learning facilitate the uptake of AI innovation and AI-prompted policy change); and
- **The readiness of wider systems** (e.g., the way that structural factors such as educational infrastructure and mechanisms of inter-organisational cooperation and multi-stakeholder coordination allow organisations and people to adopt and integrate AI innovation and AI-prompted policy change).

New funding should be considered to enable existing regulators to carry out their proposed responsibilities with respect to AI. The decentralised approach to AI regulation set out in the White Paper is designed to empower regulators to identify and assess fast-changing, context-specific risks within their regulatory remits. However, with no new funding allocated, regulators may struggle to with the significant tasks required to identify and assess AI risks, develop and enforce regulatory guidance, and continually monitor success. This could significantly impact the ability of regulators—particularly smaller regulators, or those with less expertise—to fulfil the desired role set out in the White Paper, particularly with respect to emerging technological developments.

5. What are the non-regulatory and regulatory options to address risks and capitalise on opportunities?

The diagram shared in the Introduction and Summary from the aforementioned CETaS report 'Strengthening Resilience to AI Risk' sets out a variety of policy levers, according to their aim and where they fit in the AI lifecycle. Several of these and additional policy options are presented below, according to whether they are non-regulatory or regulatory.

Non-regulatory Options

Preserving a collaborative development ecosystem. The existence of an ecosystem where people from different domains (government, industry, academia, third sector) and disciplines can collaborate to develop mutually beneficial outcomes could be helpful for the development of LLMs that support research. A proven method to enable these kinds of collaborations is through open initiatives (in an appropriate governance framework), which allows for more people to experiment, interrogate, apply, and get involved in AI, sparking innovation culture domestically and internationally. By tapping into the collective knowledge and efforts of the UK and the world's talent, open AI collaboratives allow alternative pathways to emerge for how state-of-the-art LLMs and foundation models for research or public good aims can be produced. Pre-deployment public demonstrations and deliberative processes such as citizen assemblies are especially important in flagging societal concerns before AI deployment, guiding deployment choices and designing anticipatory governance and mitigations. This should be combined with legal exemptions that allow for safety-motivated collaboration between companies building AI systems which reduces the chances of system flaws going unaddressed must also be considered to get past concerns over anti-trust regulation.

Voluntary Protocols. Government could encourage the adoption of voluntary protocols for large scale AI model safety. For example, the Partnership on AI (PAI) are developing a set of protocols for safe and responsible foundation models with representatives from industry and academia. While voluntary mechanisms are insufficient on their own, they have the advantage of potentially being more timely and agile than regulation and could be used as a basis for future regulation.

Encourage the adoption of Process-Based Governance (PBG) Frameworks. As detailed in the national public sector AI ethics and safety guidance, 'Understanding artificial intelligence ethics and safety', the PBG Framework is a governance framework that covers the design, development, and deployment process of AI and provides the foundation for effectively establishing necessary practical actions and controls, exhaustively distributing roles and responsibilities, and operationalising answerability and auditability throughout the AI lifecycle. Organising all the governance actions in a PBG Framework is a way to optimise transparency. The adoption of the PBG Framework is predicated on the commitment that, from start to finish of the AI project lifecycle, design, development, and deployment processes should be as transparent and as open to public scrutiny as possible. Greater encouragement of adopting of such frameworks and their associated transparency requirements by developers of LLMs, as well as continue to use and publicise PBGs in public sector uses of AI, would be beneficial.

AI bounties and incident sharing. The rapid pace of investment and innovation in AI means that vulnerabilities in AI systems are also likely proliferating at a quicker-than-desirable pace. In software security, bug bounty programmes have been an important mechanism in incentivising external researchers to identify and responsibly disclose risks, and policymakers could invest more resources in supporting the equivalent for AI systems. However, where initiatives like this fail to prevent system vulnerabilities from turning into harmful incidents, it is crucial that there is a systematic approach to collecting and analysing risk incidents, whether they are accidental or malicious. This could foreground patterns between incidents which would otherwise be difficult to spot if viewing them in isolation – the AI Incident Database is an existing third sector example of such a collection mechanism.

Direct engagement with industry bodies. Leading AI companies are establishing new industry bodies (see the US-based Frontier Model Forum) to oversee safe development of the most advanced models. There must be a consistent approach across UK Government to engagement with these bodies.

Coordinated watermarking and AI-enabled authorship detection. The ability to distinguish AI-generated content from human generated content is central to the production and distribution of and access to reliable information. This is of heightened importance ahead of the UK Parliamentary election next year, and both the UK Government and industry should be at the forefront of efforts to tackle this challenge, beginning by funding pilot projects to demonstrate proofs of concept.

Articulating 'red lines'. There are specific contexts where integrating LLMs into decision-making functions will be undesirable for the foreseeable future. Autonomous agents – systems which can generate a sequence of tasks that a model works on until the desired 'goal' is reached – should be a primary concern here, particularly as resources in industry and the open source community continue to pour into this space.

Regulatory Options

Mandate auditing to mitigate harms. Internal and external auditing of models are needed to understand the capabilities and limitations of generative AI systems, including LLMs. One option for regulators would be to mandate auditing to mitigate harms, as organisations would have to provide detailed information about 1) the system's development, testing, and auditing to date and 2) the developers and those responsible for internal oversight. Both of these sources of information could allow for the development of an appropriate oversight process and provide the basis for further evaluations of the model itself.

The key barriers to ensuring proper auditing processes include access to information and necessary expertise. Regarding access, what is possible from an auditing and evaluation perspective depends on what is available to the auditor. For example, API access allows for the evaluation of outputs from the end-to-end system, whereas full model access may also allow the auditor to interrogate different components within the system (given sufficient expertise and resources to do so, which could be a critical bottleneck).

While third-party auditing has advantages such as independence, given that huge sums are spent developing large scale LLMs, there is an argument that funding auditing and evaluation should predominantly fall on the developers rather than the taxpayer or civil society. One possible regulatory option is a requirement for the provision of evidence on the part of developers— i.e., the burden of proof that a system meets some specified properties. For this approach to be effective, an external body with sufficient expertise and powers to verify such claims is necessary. However, a regulatory verification model shifts some of the resource and burden of proof onto the developers themselves and does not preclude other forms of supplementary external auditing.

Mandate use of bias mitigation methodologies, impact assessments, and assurance methods. To mitigate risks and foster responsible and ethical design, development, and deployment of LLMs, the Government could mandate that companies deploying LLMs go through a series of procedures including bias mitigation, impact assessments, and assurance mechanisms. The CDEI and UK's Office for AI have highlighted the need for robust AI assurance mechanisms (including in the Government's White Paper). However, one significant barrier to AI assurance success is the lack of expertise in choosing and implementing appropriate assurance techniques. In response, CDEI have curated a collection of tools (see Portfolio of AI Assurance techniques) designed to overcome these practical obstacles and pave the way for more effective implementation. The Turing's argument-based assurance methodology is a tool worth considering, as it can generalise to considerations of fairness and transparency, for example, while remaining rooted in wider considerations such as safety and security (see Trustworthy Assurance of Digital Mental Healthcare and Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems), is another tool for consideration.

Structured communication between developers and stakeholders around the assurance process enhances the transparency and openness of LLMs. It also enables a more accessible process by allowing diverse voices to critically engage with the arguments and evidence. Central to this process is early engagement with stakeholders through participatory engagement activities, where they have hands-on influence over building assurance cases. Such an intentional, inclusive approach by regulators will enable specific needs and challenges of LLMs to be factored in while fostering greater trust amongst all stakeholders involved.

Requiring explanation-aware design and post-hoc explanations. In order to facilitate transparency and accountability, explanation-aware design and providing post-hoc explanations for models could become a requirement through regulation. Explanation-aware design ensures that systems are designed in a way that allows for accessible explanations to be given to stakeholders when AI systems are used to make decisions about them. The Turing and ICO's co-badged guidance, *Explaining decisions made with AI*, sets out best practice on how to ensure explanation-aware design throughout the entire AI project lifecycle and thereby promote transparency and accountability.

Develop Standards for the governance of LLMs. Standards are already being developed to meet the needs for assuring trustworthy AI, which can be seen in the 345 standards identified on the AI Standards Hub observatory. Standards provide organisations and regulators with an important source of knowledge for how to embed AI principles into existing processes; however, there are currently no standards that specifically address the development or governance of LLMs or foundation models more broadly. Moreover, to make the use of standards as effective as possible, there is a need to raise awareness and include a greater number of stakeholders in standards development and use, as there is limited awareness of AI standardisation across stakeholder groups, especially in SMEs and civil society. These groups also find it more challenging to engage in standards development, with roundtables that informed the development of the AI Standards Hub highlighting that the main blockers to this are 'difficulty knowing what standards are being developed', as well as financial resources and time.

Registering/licensing regime for developers. This could help to filter out the most ethically concerning AI use cases at an earlier stage, with licensing defined across dimensions like compute thresholds, algorithm design and intended use-cases. Registration – a more feasible near-term option – would involve gathering fundamental information about developers of the most sophisticated LLMs and risk assessing whether their use may violate export controls or other laws.

5a. How would such options work in practice and what are the barriers to implementing them?

Several barriers to progress must be overcome to achieve a unified approach to risks from LLMs. These include:

- **A lack of long-term, anticipatory governance functions**. Approaches to AI governance must outlast election cycles and be informed on an ongoing basis by the trajectory of research and development, as well as the views of those most directly impacted and harmed by AI systems.
- **Race-to-the-bottom dynamics between companies**. Strong leadership and direction from government is vital to ensuring that innovation is conducted responsibly and that companies are adequately addressing the impacts of AI systems on individual rights.
- **Information and skills asymmetries between government, industry and other multi**-stakeholder groups. These exacerbate race-to-the-bottom dynamics and are a significant blocker to government being able to lead the largest AI companies towards a clear regulatory agenda rather than being led by those companies.

- **Persistent tensions within the AI community**. There is a wide range of opinion and academic literature spanning long-term and current or near-term AI risks and impacts. Divisions between these groups present a confusing picture to policymakers and the public and means that policy measures to reduce harm may be stalled by indecision or be influenced by actors with the loudest voices.
- **The general-purpose nature and dual-use potential of AI**. Sector-specific regulation is a necessary but not sufficient condition of trustworthy AI. Centralised coordination across sectors will be needed for effective risk management, as discussed in our feedback to the White Paper.

In the Turing's response to both iterations of the White Paper, we outline these barriers to implementation in the context of the proposed regulatory approach and how our recommendations can be implemented in practice. We were pleased to see several of these recommendations from the first draft of the White Paper taken into consideration in the existing iteration of *A pro-innovation approach to AI regulation* White Paper, as follows:

- the inclusion in the White Paper of 'trustworthy' as an additional characteristic of the UK's overall approach to AI regulation, as well as the expansion of the principle of fairness with specific elaboration on the principle of non-discrimination citing the Equality Act of 2010 and the Human Rights Act of 1998.
- a recognition of the need for a centralised function that supports AI regulation, with reference to the Turing's Common Capacity Report commissioned by the Office for AI. The report highlighted the importance of regulators acquiring new expertise, enhancing individual, organisational, and system-level readiness for AI, and developing stronger coordination mechanisms.
- a recognition of the central role that the AI Standards Hub will play in the UK's sector-based approach to AI regulation, and as a lead partner on the AI Standards Hub, we are committed to fulfilling the UK government's ambition for the UK to be a powerhouse for responsible AI.
- the emphasis on stakeholder engagement, and the recognition that the most recent White Paper is only a step in a longer, iterative process.

In respect of the wider suite of policy levers shared in this paper's Summary, discussion of how these would work in practice is included in the full CETaS report 'Strengthening Resilience to AI Risk'.

5b. At what stage of the AI life cycle will interventions be most effective?

Interventions must take place throughout the AI lifecycle – addressing risk pathways at their source in the design and training stages, mitigating

deployment risks through implementation of clear safeguards, and redressing harmful impacts over the longer-term diffusion of AI systems across society. In the context of LLMs, any interventions must address how to govern the multi-phased character of LLMs and generative AI technologies. When foundation models are used as base models and then converted into specific generative AI applications through fine-tuning, both phases must be regulated with particular attention paid to the unique issues that arise at each stage. For example, there are several risks that occur early in the AI lifecycle which the Government's White Paper (and UK approach more generally) do not address, including poor data labelling, which can lead to harms such as unfair bias, and environmental and sustainability concerns of data storage and computational centres. Intervening only at the model deployment stage, as the White Paper proposes, limits the ability to mitigate these upstream risks. Effective AI policy should incorporate end-to-end governance approaches that address risks comprehensively, from the inception of the foundation model to the retirement of its applications.

The difficulty of predicting and mitigating the full spectrum of risks that could arise from the deployment of LLMs in different sectors means that additional governance measures focused on earlier stages of the AI lifecycle will be needed. This involves managing the way that certain models are developed and initially deployment to mitigate the full range of harms.

Post-deployment interventions are often overlooked. Large, publicly released foundation models (e.g., GPT-4) have already undergone training phases; therefore, post-training and deployment stages are the only areas in which the mitigation of future harms can take place. When viewing the current landscape of interventions, a frequently missed area of oversight is that of ongoing post-deployment evaluation. A critical component of this is change management—if a system has been evaluated at the point of deployment, how and when should it be reassessed in light of changes made to the system (including due to retraining, as well as other software updates), or in light of wider developments (such as external events that could influence its use or impact)? Post-deployment evaluation should include ongoing monitoring of system behaviour, emergent capabilities, uses, and societal impacts. Monitoring of societal impacts should include consideration of direct and indirect impacts, including broader impacts to society (for example to epistemics/misinformation, democracy, power, discrimination, employment, etc.) and impacts to different groups in society, in particular vulnerable and historically marginalised groups.

Interventions should also align with a functional model of the AI lifecycle. To understand which processes should take place at what stages, a functional model of the AI lifecycle is necessary. The model must be adaptable to different technical approaches and contexts and act as the starting point of any regulatory approach. Once established, the model will play a key role as an enabler for the identification and elucidation of common regulatory concerns and touchpoints. At the Turing, we have already developed a provisional model of the AI lifecycle (see Section 3 of

Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies), which is consistent with the CDEI's assurance roadmap. Establishing a functional model of the AI lifecycle will ensure that gaps that appear at all stages of an AI system's design, development, and deployment are covered.

5c. How can the risk of unintended consequences be addressed?

Supply chain governance could reduce unintended consequences. Policymakers must also confront the governance challenges presented by the complex and distributed supply chains that typify generative AI lifecycles, including those of LLMs. Many of these systems are comprised of parts or elements that derive from multiple suppliers, vendors, contractors, and open-source assets. This means that effective AI policy should codify end-to-end accountability and transparency mechanisms which establish a continuous chain of human responsibility across the entire project lifecycle.

Staged release. Staged release strategies can also help understand and mitigate the risks of unintended consequences of LLMs and foundation models. For example, doing an initial release to researchers, regulators, and/or auditors for a first phase of scrutiny, before a small-scale public release followed by widespread release. At each stage, there should be mechanisms in place to allow for system changes and other mitigation strategies to be put in place.

Narrowed regulations can address the higher risk of larger companies' models. The power imbalances at the ecosystem level which substantially affect the public interest should be reviewed. Because a few large private companies control the critical digital infrastructures on which the production and use of foundation models and LLM applications depend, smaller commercial and public sector companies are naturally at a disadvantage. Therefore, if AI policy aims to address such power asymmetries it will need to subject these larger entities to higher degrees of legal and regulatory intervention and control. Rectifying this problem entails the establishment of proportionate cross-regulatory processes that recognise AI as the critical infrastructure it is fast becoming, and react accordingly—for instance, by holding producers of the large base models behind AI-generated content legally responsible for their outputs and impacts.

Section 3: International context

6. How does the UK's approach compare with that of other jurisdictions, notably the EU, US and China?

The EU's AI Act would create a significantly stricter regulatory regime than the UK approach. The UK's pro-innovation approach to AI regulation, as set out in the Government's White Paper, is comparatively more flexible than the proposed EU AI

Act, which takes a risk-based approach to the regulation of AI including banning certain AI applications. In the context of LLMs, the latest version of the proposed EU AI Act creates extremely strict pre-deployment compliance requirements for LLM providers, including 'disclosing that the content was generated by AI, designing the model to prevent it from generating illegal content and publishing summaries of copyrighted data used for training' (European Parliament, 2023). Researchers have found that no current iterations of foundation models or LLMs are in compliance with the EU's requirements under the draft EU AI Act, highlighting a potentially high compliance burden (Bommasani et al., 2023). The UK has additionally delegated responsibility for AI regulation to existing regulators, including members of the Digital Regulation Co-operation Forum (Ofcom, the Competition and Markets Authority, the Information Commissioner's Office, and the Financial Conduct Authority), in contrast to the European model which aims to create a new oversight board for AI-specific regulation implementation.

Chinese regulators have proposed stringent LLM-specific regulations. Chinese regulators have specifically created a national standard for LLMs, which is substantially stricter in many areas than the proposed EU approach, and prohibits any release of generative AI that 'subverts government power and authority or questions national unity' (Coldewey, 2023). The standard also requires that:

- 'Providers assume liability and responsibility for the training data of models;
- users of the services must be verified as real people;
- personal information and reputation must be respected, or regulators may find the provider liable;
- generated content must be labelled as such;
- and [that] generative AI services will need to obtain a license to operate.' (ibid)

China has stated it is working closely with companies developing LLMs (including Chinese tech companies Tencent, Alibaba and Baidu) to ensure regulatory compliance (Jiang and Cao, 2023).

The US has emphasised voluntary compliance to reduce AI risks. The US has no current plans to specifically regulate LLMs at the federal level, in line with their overall deregulatory approach to AI, and policymakers have welcomed the creation of private sector standards setting bodies (Zakrzewski and Tiku, 2023) and voluntary commitments (see White House, 2023) on reducing the risks of foundation models and LLMs as a non-regulatory alternative. It is as yet unclear how US copyright law will apply to LLM-generated content, though US-based authors are filing a class-action lawsuit to have their content removed from LLM training (see Poritz, 2023). State-level requirements on consumer data and privacy apply to LLMs, though currently, no specific state-level legislation on LLMs has been proposed.

The UK approach aims to find a mid-point between EU and US approaches and demonstrate the viability of a sector specific approach. Overall, the UK has placed an emphasis on creating a regulatory environment which allows innovation and growth to flourish while curbing negative outcomes, as a kind of “middle of the road” approach to AI regulation. The UK’s sector-specific, decentralised approach has also encouraged different government bodies to consider AI governance as it applies to their regulatory remits. Encouraging interventions from a diversity of regulatory authorities allows each body to come up with novel solutions based on their existing governance approaches and expertise, which has, in turn, produced innovative governance initiatives (Roberts et al., 2023). The UK’s approach aims to be more flexible than the EU AI Act, reducing compliance burdens and adopting regulatory sandboxes as a mechanism to encourage AI innovation.

6a. To what extent does wider strategic international competition affect the way large language models should be regulated?

A handful of key players dominate the space. When considering LLMs through the lens of international competition, it remains clear that there are several key big players that dominate the space, specifically private companies based in the US and (to a lesser degree) China. There are concerns that the new offerings of LLMs may reinforce these large actors’ already existing market power and further concentrate AI technologies into the hands of several large firms. We have already seen antitrust cases take place prior to the explosion of LLMs, and it is possible that more cases will evolve with the expansion of generative AI models such as LLMs into workplaces and society more generally. The Federal Trade Commission (FTC) of the US commented on this, claiming that generative AI does in fact raise competition concerns. They state, ‘Generative AI depends on a set of necessary inputs. If a single company or a handful of firms control one or several of these essential inputs, they may be able to leverage their control to dampen or distort competition in generative AI markets’ (Federal Trade Commission, 2023). Additionally, the FTC outlines several areas or ‘building blocks’ of generative AI that could affect competition including, ‘a large and diverse corpus of data’ which established companies are more likely to have, in addition to ‘honed proprietary data collection tools’, labour expertise, and computational resources (see also Schrepel & Pentland, 2023). We support the CMA’s commitment to investigating competition and barriers to entry in the development of foundation models (see CMA, 2023).

Key players could distort competition through ‘bundling and tying’. The FTC warns that methods such as bundling— ‘when a company offers multiple products together as a single package,’ and tying— ‘when a firm conditions the sale of one product on the purchase of a separate product’ (Federal Trade Commission, 2023) could reduce the appeal of competitors’ offerings, especially when they only offer generative AI applications (including LLMs).

A taxonomy is necessary for assessing different types of foundation models. Foundation models differ in terms of modes of access and types of training data. In order to properly regulate this technology, it is critical that these nuances are effectively understood. Schrepel and Pentland (2023), for example, propose a taxonomy for distinguishing between foundation models. It is noteworthy also that many regulators have significantly different terminologies and approaches, and therefore, a high-level model taxonomy should be developed and discussed with sectoral regulators to translate terminology at the sectoral level.

Consider the impact of compliance costs on international competition. It is critical that regulatory compliance costs are heavily considered for small and medium-sized players in the foundation model space. As alluded to in the aforementioned points regarding access to labour expertise, market power, and computational resources, high compliance costs could remove small and medium-sized players from the market due to unaffordability. As stated by Schrepel and Pentland (2023), the 'first calls for regulation of generative AI are coming from the big players in the space who may already be showing a desire to raise barriers to entry by increasing compliance costs'.

6b. What is the likelihood of regulatory divergence? What would be its consequences?

The likelihood of regulatory divergence is high. The European Commission's publication of the world's first harmonised rules for AI in the AI Act are a reference point for any subsequent regulation, including in the UK. It remains unclear what the requirements of the EU AI Act imply for the practices adopted by UK companies seeking to do business in the EU and for UK-EU commercial relationships more generally. There is not sufficient understanding of how the UK's approach, once fully developed and implemented, will relate to the regulatory approach currently developed within the EU. Questions of divergence also arise in the context of technical standards, where they will include the issue of whether harmonised standards adopted by the EU should be adopted as designated standards in the UK.

Cross-border compliance burdens may undermine the goals of the established White Paper framework and harm innovation and trade. The White Paper signals an expectation that requirements for AI in the UK will deviate from proposed requirements in the EU; however, it is unclear whether the White Paper's proposal has been informed by an assessment of the extent to which regulatory divergence may create additional (dual-compliance) burdens for cross-border commercial relationships, potentially placing UK businesses that seek to do business in the EU at a disadvantage and stifling innovation. If compliance with EU requirements is sufficient to establish compliance with UK requirements, with the latter entailing greater flexibility/permissiveness compared to EU requirements, businesses will likely decide to comply with EU requirements by default to facilitate UK-EU commercial relationships, rather than taking advantage of regulatory flexibility in the UK. If the UK's approach is seen to be too permissive, this may affect the ability of UK businesses to participate in global markets. An incompatible UK regulatory standard might therefore lead to severe inefficiencies and burdens on business (particularly small and medium businesses) operating across borders. Asynchrony between UK and EU AI regulation and standards would also make it more difficult to prevent current and future harms linked to AI innovation.

Companies may still see the US as an ideal environment for innovation given its deregulatory approach and the high prevalence of AI resources and human capital. Despite the development of a blueprint for an 'AI Bill of Rights' and recent voluntary commitments to manage AI risk, the US does not have binding federal legislation or regulation for AI systems. The low baseline of regulatory compliance costs, combined with the greater availability of human capital in the US could still be seen by some developers as the most conducive environment in which to innovate. The UK's emphasis on a sectoral approach and the use of anticipatory regulatory approaches, such as regulatory sandboxes, may therefore provide flexibility for innovators with the certainty of cross-sectoral principles; however, it is unclear how well this approach will work to engender innovation in practice.

Regulatory divergence will likely obstruct progress in tackling global AI policy challenges. The AI ecosystem is global in nature given cross-border supply chains and consumer bases, and the harms of AI will not respect national boundaries. As emerging regulatory approaches diverge widely, change at domestic or regional level will only take individual countries so far. International cooperation and multilateral mechanisms should seek to address the global AI policy challenges, with agreed criteria for the success of policy interventions. For the UK to influence approaches to AI standards and development globally, it will need a clear vision of its role in the global AI landscape, and the appetite to expend significant time and resources to achieve ambitious targets in this area.

Future policy must recognise the mutually reinforcing relationship between domestic and global policy interventions. Being proactive in implementing the policy options described below Question 5 will put the UK in a better position to advocate for the adoption of those policies on the global stage. This will not only mitigate the worst excesses of a global divergence in AI regulation, but will generate further support and investment for the UK's domestic AI ecosystem.

References

Aitken, M., Leslie, D., Ostmann, F., Pratt, J., Margetts, H., & Dorobantu, C. (2022). Common Regulatory Capacity for AI. *The Alan Turing Institute.* https://doi.org/10.5281/zenodo.6838946

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 'Self-Consuming Generative Models Go MAD'. arXiv, 4 July 2023. https://doi.org/10.48550/arXiv.2307.01850.

Bommasani, R., Klyman K., Zhang, D., & Liang, P. (2023) Do Foundation Model Providers Comply With the Draft EU AI Act? *Stanford Center for Research on Foundation Models (CRFM)* https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Competition Markets Authority (CMA). (2023). AI Foundation Models: Initial Review. https://assets.publishing.service.gov.uk/media/64528e622f62220013a6a491/AI_Foundation_Models_-_Initial_review_.pdf

Cobbe, J., Veale, M., and Singh, J (2023, 7 April). Understanding Accountability in Algorithmic Supply Chains. *SSRN Scholarly Paper*. https://papers.ssrn.com/abstract=4430778.

Coldewey, D. (2023, 11 April) Prohibition of AI that 'subverts state power' in China may chill its nascent industry. *TechCrunch*. https://techcrunch.com/2023/04/11/prohibition-of-ai-that-subverts-state-power-in-china-may-chill-its-nascent-industry/

Creamer, E. (2023, July 5). Authors file a lawsuit against OpenAI for unlawfully 'ingesting' their books. *The Guardian.* https://www.theguardian.com/books/2023/jul/05/authors-file-a-lawsuit-against-openai-for-unlawfully-ingesting-their-books

European Parliament (2023). AI Act: a step closer to the first rules on Artificial Intelligence. https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

Federal Trade Commission (FTC). (2023, June 29). Generative AI raises competition concerns. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns

Luccioni, A. S., Viguier S., & Ligozat, A. (2022, November 3). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *arXiv Preprint.* https://arxiv.org/abs/2211.02001

Mulani, N & Whittlestone, J. (2023, June 16). Proposing a Foundation Model Information-Sharing Regime for the UK. https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk

Janjeva, A., Mulani, N., Powell, R., Whittlestone, J., & Avin, S. (2023). Strengthening Resilience to AI Risk: a guide for UK policymakers. CETaS, *The Alan Turing Institute.* https://cetas.turing.ac.uk/publications/strengthening-resilience-ai-risk

Jiang, B. & Cao, A. (2023, 7 July). China to create and implement national standard for large language models in move to regulate AI, while using its power to transform industries. *South China Morning Post*. https://www.scmp.com/tech/policy/article/3226942/china-create-and-implement-national-standard-large-language-models-move-regulate-ai-while-using-its

Perrigo, B. (2023, January 18). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *Time.* https://time.com/6247678/openai-chatgpt-kenya-workers/

Poritz, I. (2023, 29 June). OpenAI Legal Troubles Mount With Suit Over AI Training on Novels. *Bloomberg Law.* https://news.bloomberglaw.com/ip-law/openai-facing-another-copyright-suit-over-ai-training-on-novels

Rio-Chanona, Maria del, Nadzeya Laurentsyeva, and Johannes Wachs. 'Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow'. arXiv, 14 July 2023. https://doi.org/10.48550/arXiv.2307.07367.

Roberts, H., Babuta, A., Morley, J., Thomas, C., Taddeo, M. & Floridi, L. (2023). Artificial intelligence regulation in the United Kingdom: a path to good governance and global leadership?. Internet Policy Review, 12(2). https://doi.org/10.14763/2023.2.1709

Schrepel, T. & Pentland, A. (2023). Competition between AI foundation models: Dynamics and policy recommendations. *MIT Connection Science Working Paper.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4493900

Sadasivan, Vinu Sankar, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 'Can AI-Generated Text Be Reliably Detected?' arXiv, 28 June 2023. https://doi.org/10.48550/arXiv.2303.11156.

Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, Kevin M. Esvelt. (2023). ‘Can large language models democratize access to dual-use biotechnology?’ arXiv, 23 June 2023. https://doi.org/10.48550/arXiv.2306.03809

White House (2023, July 21). FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

Zakrzewski, Cat, and Nitasha Tiku (27 July 2023). AI Companies Form New Safety Body, While Congress Plays Catch-Up. *Washington Post*. https://www.washingtonpost.com/technology/2023/07/26/ai-regulation-created-google-openai-microsoft/.

Authors

The following researchers (listed alphabetically) have contributed to the response:

Fazl Barez, PhD researcher at Edinburgh Centre for Robotics, visiting PhD Scholar at University of Oxford

Professor Philip H. S. Torr, Five AI/Royal Academy of Engineering Research Chair in Computer Vision and Machine Learning, Fellow St Catherine's College, University of Oxford Department of Engineering Science

Aleksandar Petrov, DPhil Student, Autonomous Intelligent Machines and Systems CDT, University of Oxford

Dr Carolyn Ashurst, Research Fellow, Safe and Ethical AI, The Alan Turing Institute

Jennifer Ding, Tools, Practices, and Systems Senior Researcher, The Alan Turing Institute

Ardi Janjeva, Research Associate, Centre for Emerging Technology and Security (CETaS), Defence and Security Programme, The Alan Turing Institute

Dr Alexander Babuta, Director, CETaS; Director for National Security and Policy, Defence and Security Programme, The Alan Turing Institute

Morgan Briggs, Policy Research and Strategy Manager, Public Policy Programme, The Alan Turing Institute

Dr Jonathan Bright, Head of AI for Public Services and Online Safety AI Research, Public Policy Programme, The Alan Turing Institute

Stephanie Cairns, Research Assistant, Public Policy Programme, The Alan Turing Institute

Miranda Cross, Research Assistant, Public Policy Programme, The Alan Turing Institute

Professor David Leslie, Director of Ethics and Responsible Innovation Research, Public Policy Programme, The Alan Turing Institute; and Queen Mary University

Professor Helen Margetts, Director of the Public Policy Programme, The Alan Turing Institute; and the Oxford Internet Institute

Deborah Morgan, Research Assistant, Public Policy Programme, The Alan Turing Institute

Jacob Pratt, Research Associate, Public Policy Programme, The Alan Turing Institute

Vincent Straub, Research Assistant, Public Policy Programme, The Alan Turing Institute

Christopher Thomas, Research Assistant, Public Policy Programme, The Alan Turing Institute

Dr Sophie Arana, Research Application Manager, Turing Research and Innovation Cluster in Digital Twins (TRIC-DT), The Alan Turing Institute

Dr Christopher Burr, Senior Researcher in Trustworthy Systems, Head of the Innovation and Impact Hub, The Alan Turing Institute

Dr Cassandra Gould Van Praag, Senior Research Community Manager, TRIC-DT, The Alan Turing Institute

Dr Kalle Westerling, Research Application Manager, TRIC-DT, The Alan Turing Institute

Kirstie Whitaker, Programme Director, Tools, Practices and Systems Team, The Alan Turing Institute

Arielle Bennett, Programme Manager, Tools, Practices and Systems Team, The Alan Turing Institute

Malvika Sharan, Senior Researcher for Open Research, Tools, Practices and Systems Team, The Alan Turing Institute

Bastian Greshake Tzovaras, Senior Researcher for Participatory Citizen Science, Tools, Practices and Systems Team, The Alan Turing Institute

Ashley Van De Casteele, Strategic Support Officer, The Alan Turing Institute

Matt Fuller, Strategic Support Officer, The Alan Turing Institute

For comments or questions regarding this response, please contact officeofthedirector@turing.ac.uk