AI regulation: a pro-innovation approach – policy proposals

Open Consultation from: Department for Science, Innovation and Technology and Office for Artificial Intelligence

Executive Summary

Cross-Sectoral Principles (Questions 1.2 - 1.4)

- We believe that transparency requirements should be stronger for automated decisions that have a direct, material impact on people's lives or livelihoods. An acknowledgement that AI is being used to make decisions in these instances is insufficient. There should be justifications for why a decision was made to enable users to appeal or challenge these decisions.
- TAS Hub research into different explainability approaches reveals that performance can be worsened if humans have varying degrees of expertise. This means that questions around transparency and explainability might be directly linked to questions around AI education and AI literacy. https://doi.org/10.48550/arXiv.2303.12390
- Routes of redress for AI harms could be improved if they rely less on an individualised model of reporting, and are instead based on systematic, ongoing monitoring and independent assurance.

Statutory Duty (Question 2)

- We agree that a statutory duty strengthens the regulator's mandate, but it may introduce an element that makes it difficult for regulators to adaptively respond to a changing AI environment, where different types of AI will impact sectors differently. If a statutory duty on regulators goes forward, the statutory duties should be able to cater to these differences, and be put in place to ensure that other laws are upheld.

New Central Functions (Question 3)

- We think the new central functions should prioritise regulatory coordination and coherence, undertaking activities such as co-badged guidance, ongoing coordination through the Digital Regulation Cooperation Forum (DRCF), and expansion of the DRCF to include regulators relevant to AI such as the Equality and Human Rights Commission (EHRC), Medicines and Healthcare products Regulatory Agency (MHRA), among others.

Additional Education (Question 4)

- We do not agree with the planned educational outreach, as it does not go far enough. Increasing awareness of the framework will be a good thing, but educational activities for businesses and developers must facilitate understanding and application of ethical practice, and how to exercise ethical judgement, as opposed to tick box implementation.
- The TASHub has developed a systematic framework for Responsible Research and Innovation (RRI), including RRI Prompts and Cards, to assist designers and researchers with ethical product design. https://doi.org/10.1145/3064940; https://doi.org/10.1016/j.jrt.2022.100045
- Public open deliberative democratic events on the topic of AI regulation and ethics, which facilitate open conversation between the public, regulators, and developers would be an excellent way to both raise awareness and encourage an ongoing public conversation about AI and its place in our society.

Legal Responsibility (Question 5)

- We do not think existing legal frameworks are sufficient for allocating legal responsibility on topics relating to AI. Furthermore, we have concerns regarding the manner in which each regulator will apply or interpret the principles in their sector. There could be potential unfairness due to unequal treatment of parties in different sectors due to a non-standardised approach.
- Whilst the principles can be applied by regulators in a quasi-legal manner, there is scope to proactively protect society from harms eventuating from AI use in critical areas through provision of a statutory approach with rights and routes to redress. These would support the reduction of risk of harm by informing the AI industry (and those affected) of what is expected of AI product development, deployment and use. Responsibility for the outcomes of AI use must be determined prior to its adoption. We have proposed a risk-pooled shared responsibility approach: https://doi.org/10.1177/0968533220945766

Foundation Models (Question 6)

- We agree that measuring compute provides a potential tool that could be considered as part of the governance of foundation models.
- We advocate for other methods for governance beyond those proposed. A first step to the governance of foundation models would be the reinforcement of the Digital Markets Unit (DMU) as an independent regulatory body that governs and regulates foundational AI models. The DMU, with extra resources could, in effect, create and implement rules for different risks (e.g., disclosure around data being used, performance, compute) and require companies to show their work. While existing regulatory bodies are already in place and span multiple domains, these bodies are under-resourced and have failed on many occasions, especially around matters of data privacy.

Regulatory Sandbox (Question 7)

- We believe that the regulatory sandbox will not work well without sufficient expert recruitment into regulatory bodies. On the 22nd May 2023, the TASHub held an AI Regulators Workshop. Regulators mentioned that they were under-staffed when it came to technical experts on AI. The government should invest in encouraging recruitment of technical experts into regulatory bodies, upskilling staff and/or encouraging dialogues between regulators and academic experts in the fields of AI and computer science.

Response Authors:

Dr. Joshua Krook: Research Fellow in Responsible AI at the University of Southampton. He has a PhD in Law from the University of Adelaide and previously worked in technology policy for the Australian government.

Dr. John Downer: Co-I Functionality Node and Senior Lecturer in Risk and Resilience, School of Sociology, Politics and International Studies (SPAIS), University of Bristol.

Dr. Peter Winter: Research Associate in Regulation of Autonomous Systems, Functionality Node, School of Sociology, Politics and International Studies (SPAIS), University of Bristol.

Dr. Jennifer Williams: Assistant Professor in Electronics and Computer Science at the University of Southampton.

Professor Jonathan Ives: Professor of Empirical Bioethics at the University of Bristol, and Co-I TAS Functionality Node. Research interests in AI and regulating innovation in surgery.

Dr Helen Smith: Research Associate in Engineering Ethics at the University of Bristol, TAS Functionality Node. NMC Registered Nurse.

Dr. Roxana Bratu: Senior Lecturer in Public Policy, International School for Government, King's College London, UK. Research interests in AI regulatory framework, ethics and public policy.

Stephanie Sheir: Research Associate in Ethics of Autonomous Systems, University of Bristol.

Professor Robin Williams: Director, Institute for the Study of Science, Technology and Innovation and Co-I TAS Governance and Regulation Node, The University of Edinburgh.

Professor Stuart Anderson: Professor of Dependable Systems and Co-I TAS Governance and Regulation Node, The University of Edinburgh.

Dr. Phoebe Li: Reader in Law and Technology and Co-I TAS Governance and Regulation Node, University of Sussex.

Dr. Beverley Townsend: Research Associate in Law, Ethics, and Technology in the TAS Resilience Node, at the University of York.

Professor Subramanian Ramamoorthy: Professor of Robot Learning and Autonomy, and PI UKRI Research Node on TAS Governance and Regulation, University of Edinburgh.

Professor Sarvapali D. Ramchurn: Professor of AI and Director of the UKRI Trustworthy Autonomous Systems (TAS) Hub. Sarvapali has over 19 years of experience in developing AI solutions.

Citation
Trustworthy Autonomous Systems Hub. (2023) Response to: AI regulation: a pro-innovation approach – policy proposals. DOI: https://doi.org/10.5258/SOTON/P1109.

Contact details:
J.A.Krook@soton.ac.uk or John.Downer@bristol.ac.uk

About the TAS Hub:

The UKRI TAS Hub assembles a team from the Universities of Southampton, Nottingham, and King's College London. The Hub sits at the centre of the £33M Trustworthy Autonomous Systems Programme, funded by the UKRI Strategic Priorities Fund. The role of the TAS Hub is to coordinate and work with six research nodes to establish a collaborative platform for the UK to enable the development of socially beneficial autonomous systems that are both trustworthy in principle and trusted in practice by individuals, society and government. Read more about the TAS Hub here.

Q1. Our revised cross-sectoral principles, including safety and transparency

Q.1.2 Are there other measures we could require of organisations to improve AI transparency?

Transparency requirements should be much stronger for automated decisions that have a direct, material impact on people's lives or livelihoods. An acknowledgement that AI is being used to make a decision, in these instances, is insufficient. There should be further reasons and justifications for why a decision was made, including what variables and considerations the algorithm has taken into account. Without these reasons, it becomes difficult if not impossible for users to challenge decisions. Giving reasons for a decision also allows for an objective analysis of the decision by third-parties, companies and consumers, whereby systematic bias, insufficient evidence, faulty logic or other problems can be revealed, critiqued and rectified.

Automated decision-making can have a widespread impact at a pace much faster than traditional decision-making. The Australian “Robodebt Scheme” for example, involved the issuing of $1.2 billion incorrectly assigned debt notices to welfare recipients. These debt notices were based on a faulty algorithm that incorrectly calculated expected average income against Centrelink figures. Alleged fraudulent recipients were sent debt notices in error, and this continued for several years. The scheme cost the government $1.8 billion in a settlement with the affected parties. Note that this was much more costly than the alleged “benefit” of the scheme ($1.2 billion).

A similar scheme in Rotterdam, the Netherlands, automated welfare-fraud inspection. The algorithm was biassed towards those who could not speak Dutch, who were foreign nationals, or, due to these and other related criteria, showed indicators of belonging to a minority race. (Constantaras et al. 2023). The Rotterdam system triggered fraud investigations on the basis of systematic biases, resulting in minority populations being overpoliced, or living under higher levels of oversight than the general population. This could have been avoided, had the data and the variables been publicly released and/or audited by government or third-party actors.

According to *Accenture*, the creator of the Rotterdam system, the system aimed to uphold the principles of equality and diversity (Constantaras et al. 2023). Yet, despite this commitment, racial biases and discrimination were allegedly evident in the eventual application of the algorithm. A commitment to ethical principles is therefore not sufficient in and of itself to prevent bias. Auditing and testing are essential, to make sure that companies and the products they create uphold ethical and humane AI principles. Transparency is therefore not just the public release of data, but also public accountability for the management of systematic bias.

TAS Hub research into different explainability approaches also reveals that performance can be worsened if humans have varying degrees of expertise. This means that questions around transparency and explainability might be directly linked to questions around AI education and AI literacy. (Hunt et al., 2023)

Q1.3. Do you agree that current routes to contest or get redress for AI-related harms are adequate?

Current options for challenging the use of AI are often based on data protection rights. This is an individual-rights regime that relies on a tort model of litigation that serves as a disincentive and a barrier for individuals in raising attention to, and receiving compensation for, experienced harms, particularly if these individuals belong to vulnerable groups. Individuals may not be best placed to challenge the use of AI when they themselves are under-resourced, or lacking in sufficient legal representation.

As an example, the use of algorithmic systems in the workplace, typically for hiring or management, can be deployed without worker knowledge, consent, or meaningful ability to opt-

out. Workers face inherent power asymmetries in the workplace for which data protection rights provide a weak mechanism of contestation or redress. This is one of many power asymmetries that may dis-incentivize individuals from challenging AI harms.

Q1.4. How could current routes to contest or seek redress for AI-related harms be improved, if at all?

Routes of redress could be improved if they rely less on an individualised model of reporting, and are instead based on systematic, ongoing monitoring and independent assurance, across public and private sector organisations.

Measures for contestation and redress should include avenues for highly consequential risk to both the individual and to broader society, and also immediate and longer-term harms.

Q2. A statutory duty requiring regulators to have due regard to the cross-sectoral principles

Q2.7. Do you agree that introducing a statutory duty on regulators to have due regard to the principles would clarify and strengthen regulators' mandates to implement our principles while retaining a flexible approach to implementation?

Yes, a statutory duty on regulators would strengthen their mandate, but at the same time, it may introduce an element that makes it difficult for regulators to adaptively respond to a changing AI environment, especially where different types of AI will impact various sectors differently. If a statutory duty on regulators goes forward, the statutory duty should be capable of adaptation to circumstance, with a degree of flexibility and inbuilt exceptions, and be put in place to ensure that other laws are upheld.

As stated in case study 3.5, for example, the fictitious company should be required to uphold existing employment laws, and ensure that more bias and unfairness is not created from the use of AI – even better if the fictitious company can *reduce* bias and unfairness. The case study shows a potential for AI to increase transparency, which regulators should be able to recognize without a statutory duty.

Q2.8. Is there an alternative statutory intervention that would be more effective?

Statutory interventions that would be effective include requiring regulators to provide a timeline for revision, for example requiring a periodic review.

Q3. New central functions that focus on coherence across the regulatory landscape, cross-sectoral risk, and monitoring and evaluation

Q3.9. Do you agree that the functions outlined in Box 3.1 would benefit our AI regulation framework if delivered centrally?

Yes, it is important to have these central functions and activities as AI will naturally bleed across sectors. This can include the reuse and repurposing of similar underlying AI technologies (such as algorithms or data) and it may also include application areas that cut across sectors. Having aspects of the framework delivered centrally will benefit the framework.

Q3.10. What, if anything, is missing from the central functions?

Usually other countries used dedicated national labs, who have very high scientific expertise and rigour to assist with monitoring and evaluation on AI-specific technologies. As written, it's not clear what exactly M&E is doing. What constitutes monitoring, what will be evaluated, and how will it be evaluated? Are you going to incentivise regulators to follow the framework, if a statutory duty is not enacted? For example, will the regulators compete for who is achieving the best, as measured from the M&E?

Q3.11. Do you know of any existing organisations who should deliver one or more of our proposed central functions?

The Regulatory Horizons Council (RHC) is already advising regulators, but is this really working, can they provide enough? It would be better if other organisations are involved, though we do not have any suggestions of who.

Q3.12. Are there additional activities that would help businesses confidently innovate and use AI technologies?

Stakeholder engagement is mentioned, but it's not clear what shape that engagement will take. How will you ensure that industry (and others) have a voice? How will inclusivity and diversity be supported within the stakeholder group? What would greater participatory collaboration and engagement entail? It is necessary to further clarify, or provide examples of what the engagement process could look like.

Q3.12.1. If so, should these activities be delivered by government, regulators or a different organisation?

Ideally, these activities would be delivered as independently as possible, though it's clear that there are no true stakeholders outside of AI innovation, even among academics. At the same

time, academic organisations might be the most objectively placed for coordinating these activities, especially as AI regulation is a topic of interest among academics.

Q3.13. Are there additional activities that would help individuals and consumers confidently use AI technologies?

An easy-to-use reporting mechanism, similar to Yellow Card for healthcare (which MHRA uses for medicine and medical device injuries) but specifically dedicated to AI-related concerns. This should be delivered centrally.

Q3.13.1. If so, should these activities be delivered by government, regulators or a different organisation?
As mentioned above, either centrally by the government or an independent organisation capable of delivering a platform that does not deter reporting, is easy to use, and can be actioned upon where appropriate.

Q3.14. How can we avoid overlapping, duplicative or contradictory guidance on AI issued by different regulators?

One of the priorities for the central functions should be regulatory coordination and coherence, undertaking activities such as co-badged guidance, ongoing coordination through the Digital Regulation Cooperation Forum (DRCF), and expansion of the DRCF to include all regulators relevant to AI such as the Equality and Human Rights Commission (EHRC), Medicines and Healthcare products Regulatory Agency (MHRA), among others.

Q4. Additional education and awareness support for consumers, businesses, and regulators

Q4.1. Do you agree that the functions outlined in Box 3.1 would benefit our AI regulation framework if delivered centrally?

Not entirely, no. There are some problematic and simplistic assumptions - please see answers below.

Q4.2. What, if anything, is missing from the central functions?

There is a problematic focus on education and awareness about the framework and an overly simplistic assumption that increased awareness will enhance trustworthiness. Increasing awareness of the framework will be a good thing, but educational activities for businesses and developers must facilitate understanding and application of ethical practice, and how to exercise good ethical judgement, as opposed to tick box implementation.

The TASHub has developed a systematic framework for Responsible Research and Innovation, including RRI Prompts and Cards, to assist designers and researchers with ethical product design. https://doi.org/10.1145/3064940; https://doi.org/10.1016/j.jrt.2022.100045

Q4.3. Are there additional activities that would help businesses confidently innovate and use AI technologies?

As mentioned above, education with developers that focus on ethical judgement, as opposed to simply awareness of the framework, will be essential. This is to mitigate the risk of superficial tick box compliance, and encourage ethical practice.

Q4.4. If so, should these activities be delivered by government, regulators or a different organisation?

This kind of activity could be effectively delivered by teams or individuals from higher education institutions with the expertise and experience of ethics education. Prof Jonathan Ives at the University of Bristol has already been thinking about how this could be done, as part of the TAS node on Functionality.

Q4.5. Are there additional activities that would help individuals and consumers confidently use AI technologies?

Public open deliberative democratic events on the topic of AI regulation and ethics, which facilitates open conversation between publics, regulators, and developers would be an excellent way to both raise awareness and encourage an ongoing public conversation about AI and its place in our society. Increasing inclusivity and diversity - including marginalised groupings who will make use of the technology - in the discourse as active participants in the process ought to be encouraged.

Commissioning the development of educational resource packs for schools about AI and its role in society would be an excellent way to encourage thinking and early awareness of technology that will soon be ubiquitous but has significant risks.

Q4.6. If so, should these activities be delivered by government, regulators or a different organisation?

This kind of activity could be effectively delivered by teams or individuals from higher education institutions with the expertise and experience of ethics engagement. Prof Jonathan Ives at the University of Bristol has already been thinking about how this could be done, and started work on developing pilot activities, as part of the TAS node on Functionality.

Q5. The allocation of legal responsibility for AI throughout the value chain

Q5.L1. What challenges might arise when regulators apply the principles across different AI applications and systems? How could we address these challenges through our proposed AI regulatory framework?

If each regulator applies principles as per their interpretation of them, unfairness due to unequal treatment of parties in different sectors may arise due to a non-standardised approach.

Timeous and effective communication among multi-regulators would be key to coherent sectoral interpretations. The leading regulator for a specific case would be responsible for coordinating communication with other relevant regulators in order to ensure mutual supportive interpretations of the principles at issue.

Q5.L2.1 Do you agree that the implementation of our principles through existing legal frameworks will fairly and effectively allocate legal responsibility for AI across the life cycle?

No. Legal liability in England and Wales is not determined on principlism, but on legal structures - these structures could treat users unfairly. Please see Smith and Fotheringham's work here: https://doi.org/10.1177/096853322094 and here: https://doi.org/10.1177/09685332221076124

The 5 principles set out fundamental key values for AI regulation, however, huge gaps exist between these high-level principles and the practical guidelines for allocation and demarcation of legal responsibilities for actors across the life cycle. The demarcation of liability between developers and deployers would first need addressing where a third party certification could step in. In addition to tracing the liability of an individual actor, the possibility of group responsibility (such as the liability of the software development team) could be explored.

Other alternative models, such as strict liability, compulsory insurance schemes and special compensation funds (ie. The oil pollution compensation fund; the September 11th Victim compensation fund; New Zealand Compensation Scheme), for high-risk AI systems, would be beneficial when evidence is not transparent as to the responsibility of an individual or a group of actors.

Q5.L.2.2. How could it be improved, if at all?

Whilst principles can be applied by regulators in a quasi-legal manner, there is scope to proactively protect society from harms eventuating from AI use in critical areas through provision of a statutory approach with rights and routes to redress. These would support the reduction of risk of harm by informing the AI industry (and those affected) of what is expected of AI product development, deployment and use. Responsibility for the outcomes of AI use must

be determined prior to its adoption. We have proposed a risk-pooled shared responsibility approach: https://doi.org/10.1177/0968533220945766

It is useful to develop AI systems with tracing systems where chain of custody and liability could be tracked down. Particularly, at the juncture of transition from development and deployment, assurance from a third party may also be useful for examining the phased quality control. When the **ex ante** measures are adopted to prevent failure, ongoing monitoring and surveillance are still instrumental for ensuring safety in the total lifecycle. As mentioned above, additional safety measures such as insurance schemes and special compensation funds would play an integral part *ex post* to remedy the damage.

Moreover, work on developing a process to operationalise and refine high-level, abstract principles to lower-level evaluative standards for implementation by development teams has been done at the TAS Resilience hub at University of York.
https://link.springer.com/article/10.1007/s11023-022-09614-w.

Q6. Approaches to the regulation of foundation models

Q6.F1. What specific challenges will foundation models such as large language models (LLMs) or open-source models pose for regulators trying to determine legal responsibility for AI outcomes?

1. **Balance between legal responsibility and business growth:** The most important challenge in determining legal responsibility for AI outcomes is ensuring a balance between UK legal standards and business growth, as the UK Government looks to foundational AI companies as a means to jumpstart the UK's flatlining productivity. This is highlighted in the UK Government's *Pro-Innovation Approach to AI Regulation* (2023:4) which states that it wishes to "make the UK one of the top places in the world to build foundational AI companies". With this, new start-ups will enter the AI field and new foundation AI markets will emerge, centred around expectations of the commercial prospects of foundation models. Foundational AI companies are expected to bring economic prosperity through a start-up culture of innovation and entrepreneurialism. Because of this, the report plays a role in ensuring that its five new legislative principles [1] for regulating AI does not stifle AI innovation and slow the growth of businesses developing and using foundational AI models. In particular, the report emphasises the importance of a 'light touch' approach and opposes "too much responsibility" to companies developing foundational models because questions of responsibility at early stages of

innovation is likely to place “undue burdens on businesses” developing and using AI (2023: 3).

However, failing to give foundational AI companies responsibility will allow UK-based businesses the opportunity to effectively ‘interpret’ these principles in a way that presents their system or business practice in the best possible light. Given the UK Government’s push towards economic growth, their relaxed or ‘light touch’ approach is more about contributing to market share than it is about ensuring safety and effectiveness. This has led some commentators to suggest that the Government’s new approach to regulating AI is “walking a fine line between promoting innovation and protecting citizens, society and business” (Preez, 2023). It may be that a ‘light touch’ approach to regulation will move the UK economy forward but, the lack or looseness of regulatory principles and their application is likely to cloud legal responsibility for AI outcomes if something goes wrong. Such a light touch to legal responsibility at the beginning will encourage further freedom down the line to users of foundation models and impacted third parties in the AI supply chain, creating the potential for responsibility gaps to emerge. This is likely to encourage competition between foundational AI start-ups, who in support of the Government’s drive to “stimulate the UK economy” may shirk their responsibility when it comes to determining issues of accountability, liability and responsibility, allowing foundational AI companies and AI users (such as an insurance company) to be free of being legally responsible for accidents and injuries consequent on the design, engineering, or use of the model.

Put simply, a ‘light touch’ approach encourages a situation in which the responsibility and legislative requirements of foundation AI companies become lost for the sake of economic growth. It is therefore important that AI companies’ interpretation of the five principles are audited, given how interpretations will differ between individual organisations (Smit et al., 2020) or people working in them. This opens the door to investigating the professionals’ interpretations of how they apply the regulatory principles in their practices. Therefore, we encourage ‘pre-deployment’ studies to look into the tensions between practitioners’ interpretations of the regulatory principles when developing foundational AI models. A pre-deployment phase of study is vitally important in this context, given how developers and users of AI in the private sector depend on business growth and profit-making (Ada Lovelace Institute et al., 2021).

2. **Attribution of responsibility:** Because foundation models are trained on large amounts of data from various sources (e.g., 'open source'), it can be difficult to attribute responsibility for AI outcomes to a specific person or entity. This makes their behaviour difficult to predict. This attribution of responsibility becomes even more complicated when the model is 'adapted' (e.g., 'fine-tuned') by many different people or organisations for specific tasks in the AI value chain, making it even more difficult to determine who is ultimately responsible for the outcomes of these models.

   It is therefore important to note that AI users (i.e., businesses who buy these models from foundational AI companies) and then further 'adapt' these models for their own purposes opens up the risk of third-party accountability becoming diluted in the AI value chain. The 'adaptivity' of the model then, makes the process of determining legal responsibility for AI outcomes even more difficult and may allow third parties to shift the blame when negative outcomes occur.

   The challenge, then, is to ensure foundational models can be built in a way that makes clear a chain of responsibility relationships, allowing obligations to be passed from one role to another, with each link in the chain being a responsibility relationship between two roles. It should be noted that this process needs to be explicit in order to explain how the distribution of responsibilities has come about.

3. **Transparency:** The report tends to see foundation models as unproblematic, and does not mention the pitfalls of building foundational models on unlabelled data. Foundation models are trained on raw or unlabelled data, generally with 'unsupervised learning' (Merritt, 2023). According to a group of Stanford University researchers, this means no one has told the machine learning algorithm what it should be looking for (Bommasani et al., 2022). Training foundation models on unlabelled datasets and unsupervised learning is said to bring a number of opportunities to companies. For instance, unlabelled datasets saves time and money compared to traditional supervised labelling (where humans/experts are employed to manually query and re-label each data item from datasets) which is slow, expensive, and hard to govern, or reuse.

   However, training foundation models on unlabelled data also brings with it a variety of risks and uncertainties (Bommasani et al., 2022). For example, building foundation models

on unlabelled data and unsupervised learning (i.e. 'self-supervision') leads to the risk of what Snorkel AI researchers Ratner et al., (2017) call 'weak supervision' (Ratner et al., 2017) resulting in an evolving landscape of algorithmic opacity. Such opacity will make it difficult for regulators and impacted AI users and third parties to understand why these models make certain decisions or provide certain outputs (Bommasani et al., 2022), raising the challenge of transparency for companies creating foundation models for various domains. This opacity will make it difficult for regulators to evaluate the decision-making processes of foundation models and determine if they are fair, unbiased, and comply with legal requirements, and as a consequence, makes it difficult to establish levels of responsibility and liability.

In order for regulators to determine legal responsibility for AI outcomes, companies developing foundation models must ensure that rigorous methods are developed and applied to produce transparency regarding the nature of training data of foundation models. For example, Snorkel AI – a data platform company founded in 2015 – are leading the way in researching the 'weak supervision' (WS) of foundation models (Casey, 2023), resulting in over 60 peer-reviewed publications on techniques to manage weak supervision. One of these techniques includes 'Snorkel Flow', a data-centric platform that helps "make weak supervision accessible and performant" (Snorkel AI, 2023). Designed to foster public trust in foundational AI, Snorkel Flow supportw the idea that "systems need to be traceable, continuously monitored, and transparent". Looking at previous regulatory frameworks should reveal that building or using technology that fosters a transparent process is key to trustworthy AI. Regarding transparency, previous research has shown that people prefer good "everyday explanations" of AI decisions rather than technical explanations of how a decision is made (Mittelstadt et al., 2018). The importance of everyday explanations in helping AI users and public(s) to better understand how AI impacts their lives has been previously highlighted by the Government in their *Ethics, Transparency and Accountability Framework for Automated Decision-Making*.

4. **Rapid development:** The field of AI is constantly evolving, and foundation models are becoming increasingly sophisticated. Regulators are already struggling to keep up with the latest advances of foundation models – the gradual emergence of efforts to regulate the capabilities embedded in these complex technologies may come too late for some.

Foundation models such as ChatGPT – already in use across society and some domains (such as medicine and computer science) – has already raised significant questions about its potential as a tool for misinformation (Murphy, 2023; Hsu and Myers, 2023), hacking (Burgess, 2023), job loss (Berkely and Berlin, 2023; Future of Jobs Report, 2023), de-skilling (Pearson, 2023), privacy (Morrison, 2023) as well as a risk from 'bad actors' (Sparkes, 2023; Kleinman and Vallance, 2023). For example, ChatGPT has already raised questions around whether it meets data protection laws and was recently blocked in Italy with the Italian Government having already ordered OpenAI to cease collecting and processing Italian users' data until it complied with the personal data protection regulations such as GDPR by Garante Privacy (GPDP), the Italian's privacy watchdog and data protection authority (Morrison, 2023). In response, OpenAI was given 20 days to address the issues, and regulators said in mid-April that ChatGPT could return if it complied by April 30th – OpenAI did comply with the issues raised by the GPDP in late March through a series of changes, including a form to remove users' data, stricter age verification, and articles on how ChatGPT collects personal information. In the medical domain, a group of radiologists who used ChatGPT to produce a research article in the journal *Skeletal Radiology* found that it produced "false data from fictitious sources" – raising concerns of the potential harm that could come from inaccurate medical information, a problem that exacerbated by untrained readers (Murphy, 2023).

What is most concerning is that Dr. Geoffrey Hinton (a pioneer of AI), has recently joined a growing chorus of critics saying that companies developing foundational AI are "racing towards danger" and is concerned with the risks that may arise from the deliberate action of "bad actors", such as Russian President Vladimir Putin (Kleinman and Vallance, 2023). Hinton is particularly concerned about global competitiveness and how competition between foundational AI companies is forcing companies into rushed launches of models with a lack of analysis or controls to ensure responsible use. For example, some reports have claimed that ChatGPT can be used to control military drones, enabling drones to be used for target recognition (Sparkes, 2023). Such concerns recall familiar conditions of security and legal responsibility in which foundational AI models threaten regulators' capacities for innovation and responsibility. In recent weeks relating to foundational models, ChatGPT has featured as a massive security risk. Burgess (2023) insists that the "hacking of ChatGPT is just getting started", and claims people have already created prompts that "bypass OpenAI's safety systems" leading it to spout homophobic

statements, create phishing emails, and support violence - for a breakdown of ChatGPT's vulnerabilities, see Adversa (2023).

Our concern is that there are challenges in how to balance rapid development with legal management and approaches. This calls for a concerted effort to support UK foundational AI companies in shaping legal safeguards and controls alongside the development process of foundational AI. For this reason, we encourage the UK Government to move away from a light-touch approach to a more systematic, adaptive and proactive approach in responsible AI governance, such as 'compliance-by-design'[2] – applying a systematic approach to integrating regulatory requirements into manual and automated tasks and processes – an approach that has already been recommended by ForHumanity Executive Director Ryan Carrier. We must take heed of Dr. Hinton's and other experts' concerns about the dangers of foundational AI by way of ensuring each foundational AI company is: (1) enabled and empowered to conduct independent audit of AI systems through 'compliance-by-design' approaches, and (2) supported to build its own team of legal experts (bringing together, for example, human rights, data protection, transparency, accountability, competence, and equalities considerations). A positive consequence of this may stop foundational AI models going live without the company or its users considering, for example, the privacy implications of their data being used to train the algorithm. Ultimately, we stress that companies must provide documentation to prove that the model is 'safe enough' before it is released into the wild. Such documentation must include disclosures around the data that is being used, the performance of the model, and an impact assessment. Similar to existing regulatory bodies (like the FDA), foundational AI companies and AI users in each domain must conduct a safety review before deployment. This draws on the idea that there needs to be a set of safety standards for foundational AI models, for instance, a set of specific tests that a model has to pass before it gets deployed into the real world. Again, this process must draw on the expertise of independent auditors who can say that the model IS or IS NOT in compliance with standard safety thresholds.

Q6.F2. Do you agree that measuring compute provides a potential tool that could be considered as part of the governance of foundation models.

Yes. Measuring compute capacity [3] is paramount to the governance of foundation models. This is highlighted in the *A Blueprint for Building National Compute Capacity for Artificial Intelligence* report by the Organisation for Economic Co-operation and Development (OECD, 2023) which warns us that the computational capabilities required to train modern machine learning systems has multiplied by hundreds of thousands of times since 2012, due to government and private sector led initiatives within countries developing cutting-edge AI. Measuring compute is becoming increasingly prominent in governance and AI development work as those companies who are able to invest in compute can continue to reinforce socioeconomic divides, creating further differences in competitive advantage and productivity gains (OECD, 2023). For this reason, the OECD offers a blueprint to regulators and partner economies for building their own national AI compute plan along three dimensions: capacity (availability and use), effectiveness (people, policy, innovation, access), and resilience (security, sovereignty, sustainability). Additionally, measuring compute capacity can also help understand the environmental impacts of compute and better inform growing debates around climate change mitigation in order to meet climate changed targets.

In this context, measuring compute becomes a valuable tool for the governance of foundation models. By measuring the amount of compute being used to train these models, policy makers, companies, and other stakeholders could gain insight into the resource and energy requirements, as well as information about speed and performance. However, measuring the compute of foundational models presents several challenges. These include: a lack of standard metrics (there are currently no AI-specific metrics to measure the compute used to train and deploy foundational models at national or sectoral levels); complexity of model architecture; diversity of hardware; data variability; and the evolution of technology (OECD, 2023).

Q6.F3. Are there other approaches to governing foundation models that would be more effective?

The unique properties of foundation models that make them attractive as general-purpose AI may also present unknown and unpredictable risks. The issue of ‘stepping in to address risks when necessary’ may actually come too late.

A first step to the governance of foundation models would be the support and reinforcement of the Digital Markets Unit (DMU) to effectively govern and regulate foundational AI models, including predatory practices by major firms. The DMU should be encouraged to, in effect, create

and implement rules for different risks (e.g., disclosure around data being used, performance, compute) and require companies to show their work. While existing regulatory practices are already in place and span multiple domains, these bodies are under-resourced and have failed on many occasions, especially around matters of data privacy in the digital sector and Big Tech (Edwards, 2022; Garrod et al., 2023). The DMU should be supported with greater resources and research, rather than just placing funding into more projects on technology development. For example, this may take the form of pre-deployment and post deployment testing, as well as identifying/making sense of bad actors and all sorts of risky behaviour.

A second step highlights a need for the DMU to demand transparent development of foundation models. While this isn't new, the promotion of transparency in the development process is a good first step in the governance of foundation models. This would be, for example, ensuring foundational AI companies have mechanisms in place to openly share information about the model's training data, architecture, and potential biases, which opens them up to scrutiny through other actors (such as external auditors or external researchers). Doing so will help facilitate better understanding of the model.

A third step would be to establish ethical guidelines for the development and use of foundation models. Such guidelines should be specific to foundational models and address issues such as fairness, privacy, security, and the avoidance of harm.

A fourth approach would be on interdisciplinary collaboration and multistakeholder involvement in the co-production of rules and ethical principles. This could involve a diverse range of stakeholders, including researchers, policymakers, industry experts, and the general public all being involved in shaping these rules (For instance – through advisory boards, expert panels and/or public consultations).

A fifth approach would focus on robust evaluation and testing. Bommasani et al. (2022: 17), in particular, points out how foundation models challenge the existing standards of contemporary evaluation paradigms in machine learning since they are "one step removed from specific tasks". For this reason, Bommasani et al. (2022: 17) endorse the creation of three new rigorous evaluation processes to assess the performance and potential biases of foundation models. Through three central nodes of analysis, Bommasani et al., (2022: 17) emphasises: (1) a process which evaluates foundation models *directly* to measure their *inherent capabilities* as a means to inform how foundation models are trained ("intrinsic evaluation"); (2) a process which evaluates

task-specific models by *controlling for adaptation resources and access* ("extrinsic evaluation and adaptation"), and (3) a process which supports a broader *evaluation design* to provide richer context beyond measures of accuracy (e.g., robustness), fairness, efficiency, environmental impact ("evaluation design"). The creation of these three evaluation processes and testing infrastructures give hope to AI companies as they identify and mitigate biases in foundation models, especially as they look to address questions of fairness across different demographic groups. For example, adopting a process of 'intrinsic evaluation' could lead to the development and use of debiasing techniques which actively diversify the training data which, in turn, can be used to evaluate disparities in performance. In connection with that promise and against the exacerbation of unfair outcomes that arise from foundation models, Snorkel AI's data-centric platform 'Snorkel Flow' is intended as an important contribution to the identification and management of biases in inherited foundation models, with the aim of "correcting biases in AI systematically" (Team Snorkel, 2022).

A sixth approach to the governing of foundation models could be the creation of accountability mechanisms. Defining clear lines of responsibility and accountability for the development and deployment of foundation models is especially important given that foundation models (by definition) are incomplete, but can be adapted for use by an AI user (like an insurance company or telecommunications company) across different domains like industry, science, government, and academia. This could involve mechanisms for reporting and addressing any concerns or complaints raised by AI users, third parties or affected communities.

A seventh approach to governance should focus on applying updates (Dai et al., 2021) or learning such update rules (Mitchell et al., 2021). This updating and improvement of foundation model should incorporate feedback from users and stakeholders; an iterative design process which should help to ensure that the model evolves with societal needs and values.

---

[1] (1) 'Safety, security and robustness'; (2) 'Appropriate transparency and explainability'; (3) 'fairness'; (4) 'accountability and governance'; and (5) 'contestability and redress'.

[2] 'Compliance-by-design' is a process of developing a software system that implements a business process in such a way that its ability to meet specific compliance requirements is ascertained. Formal methods are typically involved to automate compliance rule verification (Kokash, 2014).

[3] AI compute capacity is defined as: “one or more stacks of hardware and software used to support specialised AI workloads and applications in an efficient manner” (OECD, 2023: 20).

Q7. An AI regulatory sandbox

On the 22nd May 2023, the TASHub held a Regulators Workshop, inviting regulators from a number of different industries to discuss the challenge of regulating AI. A consistent theme of the workshop was that regulators were under-staffed when it came to technical experts on the topic.

As a result, the government should invest in either encouraging the recruitment of technical experts into regulatory bodies, upskilling regulatory staff and/or encouraging further dialogues between regulators and academic experts in the fields of AI and computer science.

References:

GOV.UK (2023) *Pro-Innovation Approach to AI Regulation* (2023). Available Online: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#:~:text=Pro%2Dinnovation%3A%20enabling%20rather%20than,promote%20and%20encourage%20its%20uptake. [Accessed: 06/04/2023]

Adversa (2023) GPT-4 Jailbreak and Hacking via Rabbithole attack, prompt injection, content moderation Bypass and Weaponizing AI. Available Online: https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/# [Accessed: 06/04/2023]

Ada Lovelace Institute, AI Now Institute, and Open Government Partnership (2021) *Algorithmic accountability for the public sector*. https://www.opengovpartnership. org/documents/algorithmic-accountability-public-sector/

Berkely and Berlin (2023) How AI could change computing, culture and the course of history: Expect changes in the way people access knowledge, relate to knowledge and think about themselves. Available Online: https://www.economist.com/essay/2023/04/20/how-ai-could-change-computing-culture-and-the-course-of-history [Accessed: 06/05/2023].

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E. and Brynjolfsson, E. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Burgess, M (2023). The Hacking of ChatGPT is Just Getting Started. Accessed: 06/04/2023. Available Online: https://www.wired.co.uk/article/chatgpt-jailbreak-generative-ai-hacking

Casey, M. (2023) New Research expands limitations of weak supervision, foundation models. Accessed: 05/05/2023. Available online: https://snorkel.ai/new-research-expands-limitations-of-weak-supervision-foundation-models/

Constantaras, E., Geiger, G., Braun, J., Mehrota, D., Aung, H. (2023) Inside the Suspicion Machine. Wired Magazine. Available online: https://www.wired.com/story/welfare-state-algorithms/ [Accessed 02/06/2023].

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. (2021). Knowledge Neurons in Pretrained Transformers. Available online: https://aclanthology.org/2022.acl-long.581.pdf [Accessed: 05/05/2023].

Edwards, L. (2022) 'Regulating AI in Europe: Four problems and four solutions'. Available online: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf [Accessed: 18/05/2023].

Garrod, D., Pettifor, S., Meriani, M., Chinoy, K., Prota, M., Lahtinen, T (2023) 'Big tech's Growing Regulatory Burden in Europe – Failing to Prepare is Preparing to Fail'. Available online: https://www.akingump.com/en/insights/alerts/big-techs-growing-regulatory-burden-in-europefailing-to-prepare-is-preparing-to-fail#authors [Accessed: 18/05/2023].

GOV.UK. (2021). *Ethics, Transparency and Accountability Framework for Automated Decision-Making*. Available Online: https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making [Accessed: 05/05/2023].

Hsu, T., and Thompson, S.A (2023) Disinformation Researchers Raise Alarms About A.I. Chatbots. Available Online: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html [Accessed: 05/05/2023].

Hunt, W., Ryan, J., Abioye A.O., Ramchurn S.D., Soorati, M.D., Demonstrating Performance Benefits of Human-Swarm Teaming (2023) AAMAS 2023 (Winner of Best Demo Award).

Kleinman, Z. and Vallance, C (2023). AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google. BBC News. Available Online: https://www.bbc.co.uk/news/world-us-canada-65452940 [Accessed: 05/05/2023].

Kokash, N. (2014). Integrating compliance management in service- driven computing: Conceptual models and automation architecture. In R. Ramanathan and K. Raja (Eds.), *Handbook of research on architectural trends in service-driven computing* (pp. 439–480). IGI Global.

Merritt, R (2023) What are Foundation Models? Foundation models are AI neural networks trained on massive unlabelled datasets to handle a wide variety of jobs from translating text to analyzing medical images. Available online: https://blogs.nvidia.com/blog/2023/03/13/what-are-foundation-models/ [Accessed: 01/05/2023].

Mitchell, E., Lin, C., Bosselut, A., Chelsea Finn, and Manning, C.D. (2021). Fast Model Editing at Scale. In International Conference on Learning Representations. Available online: https://openreview.net/pdf?id=0DcZxeWfOPt [Accessed: 06/05/2023].

Murphy, H. (2023). '"Fictitious references" and "significant inaccuracies" could hinder ChatGPT's medical writing career.' Available online: https://healthimaging.com/topics/artificial-intelligence/chatgpts-medical-writing [Accessed: 01/05/2023].

Mittelstadt, B., Russell, C., Wachter, S (2018) Explaining Explanations in AI. Available online: https://arxiv.org/pdf/1811.01439.pdf [Accessed: 02/05/2023].

Morrison, R (2023) ChatGPT blocked in Italy over privacy concerns. Available online: https://techmonitor.ai/technology/ai-and-automation/chatgpt-blocked-italy-privacy-concerns [Accessed: 01/05/2023].

Organisation for Economic Co-operation and Development (OECD, 2023). *A Blueprint for Building National Compute Capacity for Artificial Intelligence*.

Pearson, D (2023) Generative AI: 5 concerns voiced by healthcare thought leaders. Available online: https://aiin.healthcare/topics/patient-care/digital-transformation/generative-ai-5-concerns-voiced-healthcare-thought-leaders [Accessed: 01/05/2023].

Preez, D.D (2023) 'UK sets out new approach to regulating AI that will replace 'patchwork of legal regimes'. Available online: https://diginomica.com/uk-sets-out-new-approach-regulating-ai-will-replace-patchwork-legal-regimes [Accessed: 01/05/2023].

Portillo, V., Craigon, P., Dowthwaite, L., Greenhalgh, C., Pérez-Vallejos, E. (2022). Supporting responsible research and innovation within a university-based digital research programme: Reflections from the "hoRRIzon" project. Journal of Responsible Technology, Volume 12, 100045, ISSN 2666-6596. https://doi.org/10.1016/j.jrt.2022.100045

Ratner, A., Bach, S.H., Ehrenburg, H., Fries, J., Wu, S., Ré (2017) Snorkel: Rapid Training Data Creation with Weak Supervision. Available online: https://arxiv.org/pdf/1711.10160.pdf [Accessed: 01/05/2023].

Smit, K., Zoet, M., and van Meerten, J (2020) A Review of AI Principles in Practice. In D. Vogel, K. Ning Shen & P. Shan Ling (Chairs), *Pacific Asia Conference on Information Systems (PACIS*

2021). Association for Information Systems, Atlanta, Georgia, USA https://aisel.aisnet.org/pacis2020/198

Sparkes, M (2023) Microsoft uses ChatGPT AI to control flying drones and robot arms. Available online: https://www.newscientist.com/article/2361382-microsoft-uses-chatgpt-ai-to-control-flying-drones-and-robot-arms/ [Accessed: 02/05/2023].

Snorkel AI (2023) Weak Supervision. Available online: https://snorkel.ai/weak-supervision/ [Accessed: 02/05/2023].

Townsend, B., Paterson, C., Arvind, T.T. *et al.* From Pluralistic Normative Principles to Autonomous-Agent Rules. *Minds & Machines* 32, 683–715 (2022). Available online: https://doi.org/10.1007/s11023-022-09614-w [Accessed 09/06/2023].

World Economic Forum (2023) *Future of Jobs Report*. May 2023. Available online: https://www.weforum.org/reports/the-future-of-jobs-report-2023/ [Accessed: 06/05/2023].