UK Government's Policy Proposals on

AI Regulation: A Pro-Innovation Approach

Submission 06/23

This response is provided by a working group of the Centre for Commercial Law at the University of Aberdeen. The working group is coordinated by *Dr Patricia Živković* (School of Law) and consists of *Ms Luci Carey* (School of Law), *Professor Peter Cserne* (School of Law), *Dr Claudio Lombardi* (School of Law), *Dr Michiel Poesen* (School of Law), *Dr Yaji Sripada* (School of Natural and Computing Sciences), *Dr Clare Sutherland* (School of Psychology), *Dr Robert Taylor* (School of Law), and *Tung Xuan Le* (School of Law), with comments from *Dr Titilayo Adebola* (School of Law).[*]

[*] *More information about the Centre for Commercial Law at the University of Aberdeen is available here: https://www.abdn.ac.uk/law/research/centre-for-commercial-law/.*

Table of Contents

1 EXECUTIVE SUMMARY

The analysis of our responses through ChatPDF[1] identified these ten main messages made in the document:

1. Transparency measures are an important aspect of AI regulation, but simply requiring organizations to disclose their use of AI may not be sufficient to ensure transparency.
2. The purpose of transparency measures needs to be clearly defined to provide effective and meaningful transparency.
3. We propose more robust measures for contestability and redress for AI-related harms, which would provide a clear regulatory framework, operational efficiency, and citizen satisfaction.
4. By demonstrating a commitment to citizens' rights and creating a favourable environment for businesses, the UK can position itself as an appealing and secure place to conduct AI-related activities.
5. The UK government, either independently or through international cooperation, should issue regulatory guidance or develop models, such as a Model Law or Regulation, to establish a standard for transparency in the legal domain.
6. We suggest establishing a central AI regulatory authority, with the expertise and mandate to oversee and regulate AI-related matters, ensuring consistency and specialisation in the implementation of the principles.
7. We propose that AI-based reporting of demographic differentials, such as in the case of face recognition, should be regularly tested by a state regulator to increase trust in AI systems, which would align with the objectives of the White Paper and can signal potential benefits to potential investors considering relocation to the UK.
8. We suggest that industry-based approaches, such as benchmarking algorithms through standardized tests, should accompany the context-based approach to transparency measures.
9. We suggest that more stringent AI regulation may be necessary in certain industries or sectors, such as healthcare or finance, due to the potential risks associated with AI in these areas.
10. We believe that the UK can position itself as a leader in AI regulation by taking a proactive approach to addressing potential risks and ethical concerns, while also promoting innovation and creating a favourable environment for businesses.

[1] ChatPDF is Opea AI's freely available online product that uses GPT 3.5 (https://www.chatpdf.com/) for the purpose of the textual analysis of PDFs.

2 THE RESPONSES

The revised cross-sectoral AI principles

1 Do you agree that requiring organisations to make it clear when they are using AI would adequately ensure transparency?

We welcome the general aim of the drafters to increase transparency, but we are aware that different disciplines and citizens perceive the idea of transparency differently. It is crucial to understand what transparency means, both in technological and legal terms. Transparency itself raises questions about its purpose. Knowing that AI is being used may not be sufficient in certain legal contexts, especially if there is a need to understand the inner workings of the AI system and how decisions are made. For instance, while the use of generative AI systems, such as ChatGPT, is intensely expanding in the past few months, the governmental moral duty to ask to what extent an average citizen understands the ins and outs of such a system is becoming more and more obvious. The UK citizens quickly found an everyday use for such a generative AI system, for both professional and private purposes. It is reasonable to assume that they also expect to be protected under the law from any misuse of such a freely available system in the market. Such misuse may appear in the protection of their or other people's personal data, which they share in their prompts, or from any liability that might be invoked in violation of intellectual property rights through the generation or creation based on their prompts. This is where transparency plays a crucial role in not only building the trust of citizens, but also justifying such trust.

Another key issue regards the processes that follow after transparency requirements, including i) how will transparency be utilised, and ii) which legal responsibilities will be built upon it. These are important considerations for establishing accountability, i.e. accepting responsibility for legal and ethical conduct within AI's development and deployment stream.

We also note that the question regarding "adequately ensure transparency" is currently vague. The government is asked to clarify which aspects of transparency are required, such as the purpose, usage, data collection, and audience. Clear communication to regulators or the public, depending on the context, is crucial for ensuring transparency because useful explanations are *social*. Citizens adversely impacted by AI decisions are not just looking for answers to "why AI made the decision." They are also looking for explanations that inform what the citizen could do to change the decision. This means that useful explanations are *contrastive* because they contrast the current AI (adverse) decision with other (favourable) decisions. In addition, useful explanations are *selective*, because they only present the most relevant information to the citizen instead of simply 'brain-dumping' all the internal technical details of AI. Thus, useful explanations fulfil many useful properties such as social, contrastive, and selective.

We are hereby inviting the drafters to reconsider the complexity of transparency requirements and the importance of understanding the specific aspects and goals of transparency in AI usage.

2 What other transparency measures would be appropriate, if any?

To provide adequate and effective transparency measures, one needs to define the purpose of such measures. Transparency should involve a structured approach to answering a wide range of questions and the effectiveness of transparency is measured by the system's ability to provide comprehensive answers to a broad range of questions. These answers as described above need to be social (dependent on the context) contrastive (explain what could have been done differently to reach another outcome) and selective (presenting the most relevant information).

For the purpose of selective explanations, it is critical to acknowledge that there are different types of transparency in AI, such as *global* transparency, which encompasses explaining all the decisions made by the AI model, and *local* transparency, which focuses on providing transparency for specific decisions (such as a wrong left turn made by an autonomous car).[2] Another key distinction is the difference between *partial* transparency and *full* transparency. Partial transparency, as a stepping stone toward full transparency, refers to being interested in understanding specific components or aspects of the system, even if they are not directly linked to a particular decision.[3] The community working on explainable AI believes that opening up these specific components is equivalent to opening a black box. It is seen as a way to systematically explore and answer specific questions about the system, rather than overwhelming users with an abundance of information (as explained above).

In service of achieving such balanced and useful transparency, we suggest that validated and calibrated metrics should be available to assess the quality of any transparency. For example, the National Institute of Standards and Technology runs regular vendor tests[4] to compare face recognition algorithms on standardised metrics, including suggesting clear metrics for measuring important demographic differentials across race, sex and other protected characteristics.[5] Vendors can submit their algorithms to be tested, increasing trust for clients as well as the public. Ultimately, we agree that organisations using AI should be held responsible for providing transparency. This accountability extends beyond the government or regulatory bodies, as it also involves making it clear to users and customers about the use and application of AI.

Furthermore, we note that transparency requirements should align with specific legal obligations and scientific purposes. We acknowledge the tension between predictability and explainability in the context of AI tools designed to make our lives easier. Namely, the exact moment of asking those questions and responding can have significant consequences on legal liability. We invite the drafters to reconsider the evolving nature of AI and consider the desired direction and outcomes when determining the importance of predictability. We set out three key examples here of how such measures could look:

---

[2] See more in Gesina Schwalbe and Bettina Finzel, 'A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts' [2023] Data Mining and Knowledge Discovery pt 2.3 <https://doi.org/10.1007/s10618-022-00867-8> accessed 12 June 2023.

[3] See more in Nicholas Asher, Soumya Paul and Chris Russell, 'Fair and Adequate Explanations' in Andreas Holzinger and others (eds), *Machine Learning and Knowledge Extraction* (Springer International Publishing 2021).

[4] 'Face Recognition Vendor Test (FRVT)' (*National Institute of Standards and Technology*) <https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt> accessed 9 June 2023.

[5] 'Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials' <https://doi.org/10.6028/NIST.IR.8429> accessed 9 June 2023.

1) CV sorting software,
2) Data collection, and
3) Use of AI in shipping navigation.

In the first two examples, there is a need to understand how certain decisions are made by AI systems, such as candidate selection and/or data collection and usage. While some liability claims may focus on a limited set of decisions made by AI systems, privacy claims might require a broader understanding of how data is gathered and used. Specifically, users of the AI system may need to specify the features they want the system to process to predict the desired outcome.[6] These features should be simplified for the system to effectively analyse the data.[7] It is important to note that these features, which represent different categories of data, do not necessarily have to be directly discriminatory, and in most cases, they are not. However, this does not guarantee that the system will not produce discriminatory outcomes. For instance, consider an AI system responsible for reviewing job applications and instructed to shortlist candidates who graduated from top-ranked universities; this could unintentionally result in discriminatory effects against individuals who lacked the financial means to access such educational institutions.[8] Holding AI systems accountable to the same standards as human decision-makers, such as ensuring fair and unbiased outcomes in CV scanning, is deemed essential for us and should be clearly regulated. In general, the costs of decision errors should drive the degree of transparency required.

As another key example in setting the purpose of transparency, we can also look at shipping navigation, in which predictability is crucial to ensure safety and the expectations of other users of the waterways. When it comes to liability, the ability to explain the reasoning behind AI system decisions becomes essential, particularly for collision claims where liability is determined by the tort of negligence.[9] Understanding why a system made a specific choice becomes crucial for assigning fault and resolving liability issues in collision cases. The Convention on the International Regulations for Preventing Collisions at Sea 1972 (COLREGs), given effect in the UK by the Merchant Shipping (Distress Signals and Prevention of Collisions) Regulations 1996, provide guidance to mariners on how to prevent collisions at sea but also serve as the basis for apportioning blame when collisions occur. This allocation of blame is premised on human decision-making and choices as it is permissible to depart from the COLREGs in circumstances where following them would increase danger: we suggest the decision-making process of an AI navigation system in the event of a collision requires both predictability and explainability.

Considering the various perspectives shared, we suggest that the answer to this question on transparency and liability should start by clarifying the intended goals and objectives. This would enable a more nuanced discussion about the specific measures and approaches needed

[6] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671; Frederik Zuiderveen Borgesius, 'Discrimination, Artificial Intelligence and Algorithmic Decision-Making' (Council of Europe 2018).
[7] Barocas and Selbst (n 6); Zuiderveen Borgesius (n 6).
[8] Barocas and Selbst (n 6); Zuiderveen Borgesius (n 6).
[9] See Reginald G Marsden and others, *Marsden and Gault on Collisions at Sea* (Fifteenth edition, Thomson Reuters, trading as Sweet & Maxwell 2021).

in different contexts, taking into account the desired outcomes and the potential future developments of AI.

To ensure transparency in the legal domain, we suggest that the government, either by itself or through international cooperation, or regulators (preferably a regulator specialised in AI) issue regulatory guidance or develop models, such as a Model Law or Regulation, that could establish a standard for transparency. Further specific measures would depend on specific applications but would need to be designed with a particular purpose in mind. For example, in the case of face recognition, a proposal is made that the AI-based reporting of demographic differentials is regularly tested by a state regulator. This would mean that the industry volunteers to submit its products to such testing. The already suggested context-based approach should be accompanied by industry-based approaches, such as benchmarking algorithms through standardized tests. This measure is aligned with the innovation-driven policy of the White Paper as it increases the trust of citizens in the tested AI systems.

3 Do you agree that current routes to contestability or redress for AI-related harms are adequate?

While we generally support the innovation-focused approach, and the policymakers' focus on context-based regulation, we hereby express concerns that the current approach may not sufficiently empower individuals to protect their rights if violated by AI systems and does not provide an effective means of redress. We do not consider the White Paper to be citizen-focused enough: empowering individuals and addressing their vulnerabilities in the face of AI-related harms should be a priority.

We believe that, overall, the current routes to contestability and redress for AI-related harms are not adequate, and there is a need to prioritise citizen empowerment, support, and protection in the face of potential harm caused by AI systems. At the same time, as we will explain under question 4, we do not find this to be impossible to achieve, while at the same time preserving the attraction of the UK for AI businesses and investment.

While the proposed framework states that it is pro-innovation, we are concerned that it may be emphasising the development of AI markets in the UK more than innovation itself. Although supporting the growth of AI markets is crucial for economic advancement, it is equally important to foster an environment that encourages transformative innovation. Innovation goes beyond market development and encompasses ground-breaking research, technological breakthroughs, and disruptive ideas that can lead to significant societal benefits.

To ensure that the framework truly supports innovation, it should include measures that promote transparency and contestability, research and development, and encourage collaboration between academia and industry. By nurturing innovation alongside market growth, the UK can position itself as a global leader in AI, driving advancements across various sectors and addressing complex societal challenges.

4 How could routes to contestability or redress for AI-related harms be improved, if at all?

Recognizing citizens as strong entities and constructing the regulatory framework from a citizen-centric perspective is important for raising citizens' trust in AI systems, which is rightfully recognised as one of the main goals of the drafters. The current White Paper, however, lacks a focus on empowering individuals and the citizens' role in shaping AI regulations appeared to be disregarded or suppressed. Hence, we would like to emphasise the necessity of building the regulatory framework from the ground up, centred around citizens and their empowerment. The White Paper should recognise and address the concerns and rights of individuals affected by AI systems in the future. Furthermore, it is crucial to acknowledge that citizen buy-in is vital for fostering trust in AI technologies. To achieve this, it is important to actively engage citizens in the decision-making process, promote transparency in AI development and deployment, and provide accessible avenues for public input and feedback. Involving the public should go hand in hand with the alignment of societal values, ensuring that AI technologies are meeting the needs of the majority of society and are beneficial for its members.

We emphasise here the importance of providing individuals with supporting data and evidence when they decide to contest a result produced by an AI system. Namely, knowing the number of people who have taken similar actions in the past, as well as the success or failure rates of their endeavours, can significantly empower individuals. Also, we would like to highlight the simplification of the process of lodging complaints and having someone else advocate on behalf of the individual as a preferable approach. This streamlined approach would alleviate the burden on individuals, both in terms of time and costs, by allowing them to play a more limited role once they file a complaint. We note that AI-related harms affect a broad range of consumers, employees, individuals, and businesses, and it would be more convenient and effective if they could collectively initiate legal proceedings or file complaints as a group. Furthermore, we propose that a centralized platform or regulator at the UK level is established to handle AI-related claims, similar to existing authorities for data protection.

In addition, we observe that the categorisation of consumers and businesses is made within the White Paper. While there is a specific category for consumers, we hereby raise the question of whether the distinction between consumers and businesses is entirely appropriate. Namely, businesses can encompass a wide range of entities, and the assumption that businesses are sophisticated parties while consumers are more vulnerable may not hold true in the context of AI. For instance, in a scenario where a plumber's data is utilized for AI purposes in the insurance industry, the plumber, as a business entity, may not possess the same level of understanding regarding the usage and implications of their data as a corporation would. Thus, we suggest that the binary classification of consumers and businesses should be expanded to encompass a broader range of parties.

This expanded categorisation could include entities such as charities, schools, or other organisations that may not possess the necessary knowledge or bargaining power typically associated with business-to-business contracts. By broadening the scope beyond the traditional consumer-business divide, the protection provided could be extended to a wider range of users, as the focus would shift to the importance of access to justice and information for all parties involved.

We also submit that our proposals above are aligned with the promotion of innovation in the field of AI. These proposals for more robust improvement of contestability and redress for AI-related harms would provide a clear regulatory framework, operational efficiency, and citizen satisfaction, which would make the UK an attractive destination for businesses looking to invest. By demonstrating a commitment to citizens' rights and creating a favourable environment for businesses, the UK can position itself as an appealing and secure place to conduct AI-related activities. In other words, suggested improvements align with the White Paper's objectives and they would signal the potential benefits to potential investors considering relocation.

In summary, we advocate for an inclusive approach that ensures users (both consumers and a broader range of entities) have access to justice and information. By re-evaluating the classification system and considering the diverse circumstances of different parties, the drafters can better address the needs and challenges arising from the use of AI. They should also implement collective action and centralized regulatory bodies to handle AI-related claims, which would ensure accessible and efficient redress for affected individuals.

This pragmatic approach emphasizes citizens' rights and creates a favourable business environment that could attract investment to the UK, reinforcing the mutually beneficial relationship between citizen satisfaction and economic attractiveness.

5 Do you agree that, when implemented effectively, the revised cross-sectoral principles will cover the risks posed by AI technologies?

In this regard, we would like to raise concerns about the amount of data or information that would need to be handled in AI-related cases, as they could involve substantial quantities. It is important to discuss further and decide on how to effectively inspect and process the vast amount of collected data.

Also, we note that the practical implications and challenges of implementation are not discussed in detail, and we submit that there is a need to address the complexities and practical problems that may arise. We are particularly concerned that in the absence of such discussions and proper concretisation of implementation, any implementation attempts of the White Paper will fall short of adequately addressing the practicalities.

One of such practicalities is the capacity of various regulators, such as competition authorities, privacy authorities, and consumer practices authorities, to enforce and effectively address AI-related issues should be considered. This raises questions about whether these regulatory bodies have the necessary resources and capabilities to carry out their intended roles.

In summary, we submit that the practical challenges associated with handling and processing large amounts of data in AI-related cases will usurp the effective implementation of the cross-sectoral principles. It is important firstly to deal with clarifying how accountability and the effective implementation of the mentioned principles would be achieved. Additionally, the capacity of relevant regulatory bodies to enforce and effectively address AI-related issues needs to be thoroughly considered. Any value judgement on the success of the suggested approach is premature having these practicalities in mind, and the pitfalls addressed in questions 1-5.

6 What, if anything, is missing from the revised principles?

We have no comments regarding this question.

A statutory duty to regard

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?

Introducing a statutory duty on regulators to have due regard to the principles would provide a legal framework for their actions. It would aim at the regulators prioritising and considering the principles in their decision-making processes. However, it would lack the necessary clarity and certainty in several areas. Firstly, it lacks clear guidance as to what "due regard" means in this context. "Due regard" indicates the principles must be considered, but it does not impose strict adherence to the principles themselves, and still gives discretion to the decision-maker in terms of the weight to be attached to the individual principles when making a decision.[10] Furthermore, even where a decision-maker was found to have not given "due regard" to the principles, thus breaching the statutory duty, this does not guarantee that any subsequent revised decision taking into account the principles would result in a different outcome. Consequently, the phrasing should be further refined, to provide a clearer system for accountability and legal certainty. We suggest that the wording should be redrafted into an expression with a binding effect for substantive decision-making. An example would be substituting "to have due regard" with "must follow" the principles.

Another lack of clarity stems from the lack of definition of "regulators." As it stands now, it is not clear from the White Paper which regulators would need to have due regard to the principles, and it would be necessary to define regulatory agencies to ascertain the precise focus and legal certainty that could enhance accountability and governance. Clarification is needed as to whether regulators refer to specific executive or public bodies, or indeed all of them. The desirable clarification would confirm the applicability, thus helping with legal certainty and accountability.

It is important to note here that in connection with the first argument, the spectrum of different executive agencies and public service bodies could contribute to the lack of harmonised standards of application of these principles. The application of the principles may vary depending on the context and purpose of AI, and a one-size-fits-all approach might not be appropriate for all regulators.

Additionally, regulators often rely on third-party private systems for various functions, which may involve contractual or licensing arrangements. The liability and accountability of these private actors in the development and use of AI systems should be a significant consideration. Although the government's primary focus is on ensuring the effectiveness of its agencies, the broader question of private actor liability should not be neglected. The proposed approach would not adequately address this issue.

[10] See *R (on the application of Friends of the Earth Ltd and others) (Respondents) v Heathrow Airport Ltd (Appellant) [2020] UKSC 52*.

It should also not be left unclear whether any statutory duty to have “due regard” to the principles applies only to decisions pertaining to the adoption of AI technology, or whether it extends to any other or indeed all decisions that a regulator can make. Clarity on this point is needed as it will affect the types of decisions potentially open to legal challenge for alleged breach of any statutory duty.

Another important aspect to consider is how regulators' adherence to the principles will be monitored and reported, as there should be a feedback policy in place to ensure transparency and accountability. If the current proposal is implemented, such reports should be made available to Parliament for review, which would allow for scrutiny and oversight.

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

While it is challenging to determine a singular, universally applicable alternative, there are potential options worth exploring. Some of these alternatives could include:

1. Establishing a Central AI Regulatory Authority: Instead of placing the burden on individual regulators, a dedicated central authority focused on AI governance could be created. This authority would have the expertise and mandate to oversee and regulate AI-related matters, ensuring consistency and specialisation in the implementation of the principles. Whilst we acknowledge that the use of AI may depend on context, there are significant similarities in terms of market complexity and the need to make informed decisions on what to use and how to implement the principles and rules over time.

Furthermore, although such an authority would operate independently within its statutory obligations, it should nevertheless be accountable to both the Government and Parliament, thus ensuring proper scrutiny and oversight and establishing a system of checks and balances. This approach would alleviate liability concerns, as government entities using approved AI systems could rely on the regulatory agency's assessment and approval processes.

2. Developing and Implementing Sector-Specific AI Regulations: Rather than a broad statutory duty, sector-specific regulations could be introduced to address the unique challenges and considerations associated with different industries. Regulators create guidelines and codes of practice, which can be submitted to the government for approval. This approach would allow for more targeted and contextually relevant interventions, promoting effective regulation while accommodating the specific needs and dynamics of each sector. While this may serve as an alternative to statutory intervention, it is important to note that non-statutory frameworks would not necessarily prevent legal action. If a regulator blatantly disregards the codes of practice and procedures it has implemented, individuals could argue that their procedural rights or legitimate expectations have been violated, leading to potential legal challenges. However, the legal implications may not be as straightforward as with a statutory duty.

3. Enhancing Collaboration and Coordination Among Regulators: A collaborative approach involving multiple regulators working together could be adopted, in addition to statutory duty and the other two measures mentioned above. This would encourage information-sharing, coordination, and joint efforts in addressing AI governance issues. By leveraging the expertise and resources of different regulatory bodies, a more comprehensive and cohesive regulatory framework could be established.

New central functions to support the framework

9 Do you agree that the functions outlined in section 3.3.1 would benefit our AI regulation framework if delivered centrally?

We have no comments regarding this question.

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

We have no comments regarding this question.

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

We have no comments regarding this question.

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

The AI market, although nascent, is already highly concentrated, and leaving it largely unregulated will only further its concentration ratio.

One example highlighted in the public consultation document (p.14) is the application of AI in the agricultural sector, which raises concerns about the already high concentration of this market. Without adequate regulation, the integration of AI technologies into agriculture could further exacerbate market concentration. It is crucial to recognise that concentrated markets may hinder competition, stifle innovation, and limit the choices available to farmers and consumers.

To address this issue, the proposed framework should incorporate regulatory measures that promote fair competition, prevent monopolistic practices, and encourage a level playing field. Implementing safeguards, such as transparency requirements, data access provisions, and antitrust regulations, can help mitigate the risks of market concentration and ensure that the benefits of AI in agriculture are widely accessible.

User/Citizen-Centred Approach

Shifting the focus to a user/citizen-centred approach is paramount to fostering an AI market that truly serves the needs and interests of individuals. Placing citizens at the centre of AI development ensures that technology is designed, deployed, and regulated in a manner that respects their rights, values, and well-being. By prioritizing user-centricity, the UK can create an inclusive and responsive AI ecosystem that benefits all stakeholders.

Adopting a user/citizen-centred approach ensures that AI technologies are designed and deployed to meet the specific needs and preferences of users and citizens. This approach involves actively seeking user feedback, engaging in public consultations, and involving diverse stakeholders in decision-making processes. By placing users at the centre, the UK can create an AI market that focuses on enhancing user experiences, respecting privacy rights, and addressing societal concerns.

Growth of AI Smaller Competitors:

A user/citizen-centred approach encourages the growth of smaller AI competitors. When the development of AI technologies is driven by user needs, smaller companies have an opportunity to thrive by offering specialized, user-centric solutions. These smaller competitors often possess the agility to adapt quickly to changing user requirements and can introduce innovative approaches that address specific user pain points. Promoting a diverse and competitive AI market benefits users through increased choice, improved quality, and greater innovation.

Development of Related Markets:

A user/citizen-centred approach to AI can also stimulate the growth of related markets, such as legal and auditing. As AI technologies become more prevalent, there is a need for legal counsel, regulations, and auditing mechanisms to ensure their responsible and ethical use. By focusing on the citizen's perspective, the UK can foster the development of legal and auditing services specialised in AI-related issues. This creates opportunities for legal professionals, auditors, and experts to support the fair and accountable deployment of AI technologies while protecting the rights and interests of citizens.

Protection of Citizens and Nudging Innovation:

A user/citizen-centred approach provides a framework for protecting citizens' rights and interests in the AI landscape. By actively involving citizens in decision-making processes and prioritising their well-being, the UK can establish regulations and guidelines that safeguard against the potential risks associated with AI, such as bias, discrimination, or privacy breaches. Furthermore, a user/citizen-centred approach nudges innovation in the right direction by encouraging the development of AI solutions that align with societal values, respect ethical considerations, and promote the public good.

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

We have no comments regarding this question.

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

We have no comments regarding this question.

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

We have no comments regarding this question.

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

We have no comments regarding this question.

Monitoring and evaluation of the framework

15 Do you agree with our overall approach to monitoring and evaluation?

We have no comments regarding this question.

16 What is the best way to measure the impact of our framework?

We have no comments regarding this question.

17 Do you agree that our approach strikes the right balance between supporting AI innovation; addressing known, prioritised risks; and future-proofing the AI regulation framework?

We have no comments regarding this question.

18 Do you agree that regulators are best placed to apply the principles and government is best placed to provide oversight and deliver central functions?

We have no comments regarding this question.

Regulator capabilities

19 As a regulator, what support would you need in order to apply the principles in a proportionate and pro-innovation way?

We have no comments regarding this question.

20 Do you agree that a pooled team of AI experts would be the most effective way to address capability gaps and help regulators apply the principles?

Tools for trustworthy AI

21 Which non-regulatory tools for trustworthy AI would most help organisations to embed the AI regulation principles into existing business processes?

The current trend in AI is to reuse open-source tools and libraries to the maximum extent, building custom tools and libraries only to fill gaps in open-source resources. This resulted in most AI organisations using standard MLOps technology stacks to run AI lifecycle projects. These MLOps technologies help to bring all the data and metadata related to an AI project into one place. For example, tools such as Neptune.ai offer such functionality (Neptune.ai is used here only as an example of desired functionality than endorsing the product/company/services).[11] Transparency requirements and their associated metrics could be made available as part of the MLOps stack so that AI organisations could meet the requirements as part of their regular AI development projects. In addition, AI organisations could be issued certifications (like ISO 9000 family) to help public and private businesses confidently engage with AI organisations under full disclosure of their capabilities.

[11] 'Neptune.Ai' <https://neptune.ai/> accessed 13 June 2023.

Certification for this purpose could be a leading practice globally, with the UK pioneering such a system. It would not only foster and enforce cross-sectoral principles as introduced by the White Paper but could also place the UK as the world leader in global efforts of AI certification.

Final thoughts

22 Do you have any other thoughts on our overall approach? Please include any missed opportunities, flaws, and gaps in our framework.

While we support the cross-sectional approach to regulating AI systems presented in this White Paper, some concerns need to be addressed:

Firstly, there is an issue of the discrepancies in AI regulation across different disciplines, such as the different approaches to accessing data for medical research and collecting online activity through cookies, without clear explanations for these differences.

Secondly, we believe a thorough institutional design is necessary to ensure the consistent application of AI principles and their integration within existing and future frameworks.

Thirdly, the lack of clarity and specificity in future legislative and regulatory efforts in this area is a concern for all disciplines involved and citizens' safety.

Finally, there is a need for greater institutional cooperation and the creation of a platform to facilitate it.

Legal responsibility for AI

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?

We have no comments regarding this question.

L2.i. 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?

We have no comments regarding this question.

L.2.ii. How could it be improved, if at all?

We have no comments regarding this question.

L3. If you are a business that develops, uses, or sells AI, how do you currently manage AI risk including through the wider supply chain? How could government support effective AI-related risk management?

We have no comments regarding this question.

Foundation models and the regulatory framework

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

Because foundation models can be exploited in multiple ways in downstream applications, it is important to offer guidance and support for disclosing known vulnerabilities of the downstream applications. It is a standard practice recently to disclose data related to the model runtime behaviour as part of model cards.[12] As stated below, AI organisations could be encouraged to collect and then disclose a broad range of data/meta-data and performance statistics as part of their MLOps technology stack (see our response to Question 21 above).

From a legal standpoint, the primary hurdle for regulators is determining the allocation of legal responsibility for the outcomes of AI. Specifically, addressing the issue of systemic risk poses a significant challenge. In addition to the aforementioned reporting obligations, an alternative solution could involve the licensing of LLMs, placing the responsibility on the licensors. In such cases, licensors would undergo audits, and enforcement efforts would be concentrated on ensuring compliance with these obligations. However, this approach may introduce potential anticompetitive risks, particularly concerning data access and the broader discussion surrounding public investment and private innovation.

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

No, monitoring the amount of computing power used to train foundation models is not entirely satisfactory. This is certainly useful in ensuring the green credentials of foundation models. Foundation models are increasingly viewed as general-purpose computers that execute controlled natural language prompts (programs).[13] AI organisations building applications using foundation models could be asked to disclose the subset of prompt programming exploited in their applications and what have been the issues observed while prompt programming in their lab tests. Although not possible with every new application, with time adversarial prompts could be made available as part of standard MLOps tools (see our response to question 21 above) for AI organisations to use as useful examples to design their own internal lab tests.

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

Foundation models are versatile and can be programmed (using natural language-based prompts/instructions) to perform a broad range of tasks. While newer applications would continually be innovated, it is expected that a select few capabilities of the foundation models are most likely to be exploited in most of these innovations. In this case, it should be possible

[12] See Margaret Mitchell and others, 'Model Cards for Model Reporting', *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2019) <https://doi.org/10.1145/3287560.3287596> accessed 12 June 2023.

[13] See Yongchao Zhou and others, 'Large Language Models Are Human-Level Prompt Engineers' (*arXiv.org*, 3 November 2022) <https://arxiv.org/abs/2211.01910v2> accessed 13 June 2023.

to encourage organisations and AI researchers to create adversarial prompts to test new applications. Lists of these adversarial prompts could be made available as part of MLOps stacks so that AI organisations could carry out in-house testing and disclose the testing results as part of model cards which is an accepted practice to publish key information about ML models.

From the legal stance, the implementation of this practice should be a regulatory requirement before the deployment of the AI product in the market, which can then be regularly tested by the central AI regulator. We believe that such a regulatory practice is essential to safeguard against algorithmic discrimination, biases, and malicious behaviour. By implementing these measures, we can proactively mitigate potential risks associated with AI technologies and ensure ethical and responsible use.

AI sandboxes and testbeds

S1. Which of the sandbox models described in section 3.3.4 would be most likely to support innovation?

We agree with the decision to start with a single sector and multiple regulator sandbox because this strategy guarantees to uncover the issues arising from AI organisations fulfilling multiple regulations. When a high-value sector such as healthcare (please see our response to S4 below) is selected to support AI organisations building applications based on foundation models for the healthcare industry there are great opportunities to investigate interactions among different regulatory regimes such as information governance surrounding applying AI to unconsented patient data and Patient and Public Involvement and Engagement (PPIE).

S2. What could government do to maximise the benefit of sandboxes to AI innovators?

Although government currently plans to offer a sandbox model that offers customised advice and support, we suggest (see our response to F1-3 above) government to work with AI organisations and researchers to build resources (e.g., adversarial inputs) to test AI applications by AI organisations in-house. Because most organisations use one of the standard MLOps frameworks to build AI, these resources could be made available from within the standard MLOps frameworks to make it easy for AI organisations to join sandbox programs.

S3. What could government do to facilitate participation in an AI regulatory sandbox?

As stated above, the best strategy to facilitate participation in sandbox programs is to add regulatory resources and services into standard MLOps frameworks so that regulatory work becomes part and parcel of AI development rather than an afterthought.

S4. Which industry sectors or classes of product would most benefit from an AI sandbox?

We recommend two high-value sectors where the authors of this document are working on introducing AI applications (derived from foundation models), healthcare and the Nuclear Decommissioning Authority (NDA). Both these sectors are highly regulated already. Information governance regulation and PPIE currently define guidance to AI in the healthcare sector. Trusted Research Environments (TRE) such as the Grampian DaSH (Data Safe Haven) currently offer support for AI research in healthcare where standard workflows for the entire

AI lifecycle exist. Starting with a sandbox offering custom advice to TREs to scale up AI innovation would have huge societal benefits and offer opportunities to understand the issues involved in applying safe AI without breaching patient privacy guidelines. Likewise, NDA requires multiple forms of AI because parts of NDA activity must use AI and Robotics to keep humans away from hazardous material. Currently, highly regulated workflows exist to govern activities related to handling nuclear waste where great opportunities exist to study human-AI collaborations.

3 BIBLIOGRAPHY

Asher N, Paul S and Russell C, ‘Fair and Adequate Explanations’ in Andreas Holzinger and others (eds), *Machine Learning and Knowledge Extraction* (Springer International Publishing 2021)

Barocas S and Selbst AD, ‘Big Data’s Disparate Impact’ (2016) 104 California Law Review 671

‘Face Recognition Vendor Test (FRVT)’ (*National Institute of Standards and Technology*) <https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt> accessed 9 June 2023

‘Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials’ <https://doi.org/10.6028/NIST.IR.8429> accessed 9 June 2023

Marsden RG and others, *Marsden and Gault on Collisions at Sea* (Fifteenth edition, Thomson Reuters, trading as Sweet & Maxwell 2021)

Mitchell M and others, ‘Model Cards for Model Reporting’, *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2019) <https://doi.org/10.1145/3287560.3287596> accessed 12 June 2023

‘Neptune.Ai’ <https://neptune.ai/> accessed 13 June 2023

Schwalbe G and Finzel B, ‘A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts’ [2023] Data Mining and Knowledge Discovery <https://doi.org/10.1007/s10618-022-00867-8> accessed 12 June 2023

Zhou Y and others, ‘Large Language Models Are Human-Level Prompt Engineers’ (*arXiv.org*, 3 November 2022) <https://arxiv.org/abs/2211.01910v2> accessed 13 June 2023

Zuiderveen Borgesius F, ‘Discrimination, Artificial Intelligence and Algorithmic Decision-Making’ (Council of Europe 2018)

R (on the application of Friends of the Earth Ltd and others) (Respondents) v Heathrow Airport Ltd (Appellant) [2020] UKSC 52