AGENCY RESEARCH

AGENCYRESEARCH.NET

REIMAGINING AI GOVERNANCE

A Response by AGENCY to the UK Government's White Paper AI Regulation: A Pro-Innovation Approach.

Rebecca Owens, Jehana Copilah-Ali, Maxim Kolomeets, Shrikant Malviya, Karolina Markeviciute, Shola Olabode, Tasos Spiliotopoulos, Han Wu, Viana Nijia Zhang, Kovila Coopamootoo, Abigail Durrant, Karen Elliott, Stamos Katsigiannis, Cristina Neesham, Lei Shi and Ben Farrand

June 2023

About Us

We are a diverse and multidisciplinary team of academics with expertise in computer science (security, AI, gamification, HCI), social sciences, business, economics, law, and media studies. Our members come from esteemed academic institutions, including Newcastle University, Durham University, University of Birmingham, King's College London, Royal Holloway University of London and University of Surrey.

Our research is supported by UK Research and Innovation through the Strategic Priority Fund as part of the Protecting Citizens Online programme. Grant title: AGENCY: Assuring Citizen Agency in a World with Complex Online Harms. Grant reference: EP/W032481/2.

Our Mission

The online world is, for many people, a curious but uncertain one. It enriches many facets of life but, at the same time, exposes citizens to a variety of threats that may cause harm to them, their loved ones and wider society. There is growing evidence that many such harms result from a complex interaction of societal processes driven by diverse stakeholders. When these complex harms happen to citizens, they are not purposely caused or easily controlled. The AGENCY[1] project is motivated by the firm belief that establishing citizen agency is a sine qua non for any transformative approaches that reduce these complex harms. Citizens need to be empowered through technologies and user-centred tools that enable them to gain a sense of control, ownership, security, trust, and assurance in their online activities.

AGENCY aims to establish interdisciplinary co-design principles, technology foundations and collaborative governance procedures to assure online citizen agency in the presence of multiple stakeholder interests. To do this, it utilises different technology case studies and a multiple-stakeholder approach to provide a profound understanding of the role of agency in reducing complex online harm. It delivers collaborative methods, technological building blocks and scientifically grounded best practices for our society to provide more proactive and structured approaches to protecting citizens online.

1 EP/W032481/2.

1. Abstract

On March 29, 2023, the UK Government released a white paper outlining its plans to implement a pro-innovation approach to Artificial Intelligence (AI) regulation and strengthen the UK's position as a global leader in AI.

As part of the white paper, the government has developed five key principles to guide regulators. These principles encompass safety, security and robustness, appropriate transparency and explainability, fairness, accountability and governance, and contestability and redress. To ensure an effective framework, a flexible sector-specific approach is combined with a central function that coordinates, monitors, and adapts the overall framework. This approach aims "to drive growth and prosperity by boosting innovation, investment, and public trust in AI", allowing for the opportunities posed by AI to be realised.

This response by members of the multi-disciplinary AGENCY team provides comments on the white paper and details our recommendations on how the framework can be improved to ensure that AI is developed in an ethical and socially responsible way that protects citizens against complex online harms.

2. Overview

Given our multi-disciplinary expertise AGENCY welcomes the opportunity to provide comments on the White Paper 'Establishing a Pro-innovation Approach to Regulating AI' and details below a summary of our position on the proposed regulations. Our answers to the Government survey are enclosed within the Appendix.

- **The AI Governance framework in the UK is currently underdeveloped. AGENCY supports the Government's mission to reconfigure the AI governance landscape and make the UK a global leader in AI.**
- **Due to the dynamic pace of innovation within the AI field, AGENCY urges the swift implementation of the UK's governance framework and the provision of extra resources to regulators to enable them to deal with the complex harms AI may pose.**
- **Aligning the UK regulatory framework with international standards is crucial to promoting innovation. Inconsistencies in regulations can create contradictory guidance and impede progress. Therefore, prioritising regulatory alignment facilitates a conducive environment for AI innovation to thrive.**
- **To ensure user confidence within new technology AGENCY endorses the addition of "trust/trustworthiness" as a guiding regulatory principle.**
- **Transparency plays a pivotal role in addressing concerns surrounding AI. It is essential to establish clear and open channels of communication between stakeholders.**
- **Education is essential for empowering individuals to effectively engage with AI. Key elements in this process include workshops, tutorials, and integration into the national curriculum and development of a holistic national effort. Through these initiatives, individuals gain a comprehensive understanding of AI concepts, applications, and implications, enabling them to make informed decisions and navigate emerging technologies.**

- **User autonomy should be prioritised in the design of AI systems, requiring additional information to identify disparities and ensure equitable access. Presently, neural network AI systems operate in a manner that even their creators cannot confidently decipher, making it crucial to incorporate a well-designed interface that empowers users with control over the AI's behaviour, preferences, and data usage. To achieve this, an intuitive and user-friendly interface should be implemented, allowing users to navigate and personalise their AI experience easily. Moreover, it is essential to develop an accessible interface that caters to individuals with diverse backgrounds and abilities, thereby further enhancing user autonomy. Therefore, AGENCY recommends that regulatory bodies should support investment and activities in unpacking what it means to provide control and autonomy.**

- **To enhance transparency and enable users to make informed decisions about AI systems, it is important to adopt regulatory measures similar to the EU's proposed 'right to know.' These measures ensure that users are informed when they are engaging with a chatbot rather than a human, allowing them to discern the nature of their interactions and make informed choices based on that knowledge.**

- **Regulatory sandboxes allow innovators and regulators to understand the challenges of bringing innovative AI products to market. We believe they could benefit a range of sectors such as financial services, communications and healthcare, but they must be supported by periodic evaluations which provide timely and transparent feedback to participants and other stakeholders.**

- **We support the consistent assessment and monitoring of new AI technology by the 'central functions' but believe that this should be complemented by mandatory reporting by the AI research and development community and a statutory power for regulators to perform audits of AI innovators. This should involve the disclosure of data that is used to train the AI to regulators such as the ICO. In addition, all AI innovators should be required to conduct a mandatory ex ante impact assessment akin to a 'stress test' to anticipate any complex harms that may develop before the AI technology is deployed in public domain. This should be supplemented by the adoption of an adaptable CDR framework to assess impact of data and digital technologies across the whole process from design to delivery into the digital world.**

- **AI risks should be managed by the creation of a clear and comprehensive AI strategy tailored for the organisation in question, which should include goals, timelines, resource allocation, risk management, a roadmap for implementation as well as ethical considerations and guidelines for responsible AI use, enabling to optimise AI adoption and implementation, and concurrently identify the emerging risks. For instance, certain risks could be managed by ensuring appropriate AI infrastructure for the vast amounts of data employed, educating and training employees, investing in talent and multidisciplinary expertise acquisition, collaborating with other institutions, seeking feedback, staying informed about technological, regulatory AI developments, monitoring and evaluating AI performance.**

- **Whilst the overall approach does a good job of explaining "what" needs to be done under the five principles we recommend organisations engage with the Corporate Digital Responsibility (CDR) framework (https://corporatedigitalresponsibility.net/) as a means to incorporate the five principles long-term. This will allow the creation of an organisational culture that encourages ethical practices around AI in ways that are socially, economically, and environmentally responsible, thus considering the impact on future generations in a way that ensures responsible innovation through adherence to ethical procedures that do not hinder innovation. One such approach may be the adoption of the UKRI Framework for responsible AI research and innovation (https://www.orbit-rri.org/about/about-rri/).**

- **The impact of the new regulatory framework should be constantly measured through quantitative and evidence-based risk assessment dependent on "real-world" data such as historical incident reports and use cases and stakeholders' feedback and surveys. Findings should be communicated clearly to relevant stakeholders to inform future policy development.**

- **AI is currently embedded in a lot of people's daily lives. National campaigns are needed that integrate (1) stakeholders from different walks of life, (2) companies to model videos for public consumption explaining how they use AI in delivering products and services, (3) broadcasting/news industry.**

3. Conclusion

AGENCY supports the government's vision to make the UK a global leader in AI innovation. We recognise the value of a principles-based approach that allows flexibility in accommodating future advancements. However, this should be complemented by the allocation of additional resources to interdisciplinary AI research and development, ensuring that the UK remains at the forefront of responsible AI innovation.

We strongly endorse the adoption of a multi-stakeholder approach similar to the one embraced by the AGENCY project. By bringing together diverse perspectives and expertise, we can ensure that AI systems are developed in an inclusive, ethical manner that empowers users. This should be reinforced by a legal obligation for AI innovators to periodically report their progress to regulators and a mandatory ex-ante impact assessment to understand any risks and complex harms before introducing AI systems into the public sphere. The adoption of a CDR framework should complement this allowing the creation of an organisational culture that promotes ethical practices in AI. Through this, we can construct an environment that supports responsible practices and fosters trust among the public creating a social, economic, and environmentally sensitive arena for the UK to become a leader in AI innovation.

4. Appendix

Q1. Do you agree that requiring organisations to make it clear when they are using AI would improve transparency?
A. Strongly agree.

Q2.Are there other measures we could require of organisations to improve transparency for AI?
A.The incorporation of a Corporate Digital Responsibility framework towards creating an organisational culture that encourages ethical practices around AI in ways that are socially, economically, and environmentally responsible, thus considering the impact on future generations in a way that ensures responsible innovation through adherence to ethical procedures that do not hinder innovation.

Q3. Do you agree that current routes to contest or get redress for AI-related harms are adequate?
A. Somewhat disagree.

Q.4 How could current routes to contest or seek redress for AI-related harms be improved, if at all?
A. Individuals seeking compensation for damages resulting from AI failures would typically have to pursue negligence claims, requiring them to demonstrate that the defendant had a duty of care, breached that duty, and caused an injury. However, establishing a causal link becomes challenging when the AI system's behaviours are unforeseeable and it operates independently. As a result, an approach similar to the EU's proposed AI Liability Directive could also be explored. This would make it easier for users suffering AI harms to bring civil liability claims against manufacturers and organisations using AI by creating a rebuttable presumption of causality and allowing users to be protected.

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

Q6. What, if anything, is missing from the revised principles?

A. Although building public trust is mentioned throughout the document, 'Trustworthiness/Trust' does not explicitly feature as a principle. Through promoting the development of 'trustworthy AI' user confidence and engagement is increased. This can be operationalised with CDR as Trust is the first CDR principle see: https://corporatedigitalresponsibility.net/cdr-manifesto-english.

Q7: 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?

A. Strongly agree.

Q8. Is there an alternative statutory intervention that would be more effective?

A. This could be accompanied by a prescriptive requirement for regulators to audit AI developers regularly and the creation of a statutory requirement that developers maintain consistent communication with regulators, keeping them informed about their progress. This would be complemented by a mandatory ex ante impact assessment to understand any risks and complex harms before introducing AI systems into the public sphere.

Q9. Do you agree that the functions outlined in section 3.3.1 would benefit our AI regulation framework if delivered centrally?

A. Monitoring and evaluating the framework as a whole: strongly agree.
Assessing and monitoring cross-economy risks arising from the use of AI: strongly agree.
Scanning for future trends and analysing knowledge gaps to inform our response to emerging AI: strongly agree.
Supporting AI innovators to get new technologies to market: strongly agree.
Promoting international alignment on AI regulation: strongly agree.

Q10. What, if anything, is missing from the central functions?

A. The proposals need a global approach which brings together stakeholders in government, civil society and the tech industry. This could be through workshops and forums organised by regulators, allowing key stakeholders to come together and engage in dialogue.

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

A. Organisations such as CyberNorth and the Open Data Institute may provide pragmatic insight into the utilisation of AI. Sector specific stakeholders may also be useful such as UKRI given their insight into specific issues.

Q12. Are there additional activities that would help businesses confidently innovate and use AI technologies?

A. Yes. To understand users' needs and preferences comprehensively, it is crucial to recognise the intended and unintended consequences of AI technology on people's lives. Therefore, businesses need to address any resulting disadvantages or complex harms proactively. One way to achieve this is by adopting user-centric design principles during the development of AI applications. This involves actively involving end-users by soliciting feedback, conducting user testing, and implementing iterative improvements based on their input.

Furthermore, companies should consider external audits and seek third-party validation to ensure the reliability, fairness, and compliance of AI systems and algorithms. These independent assessments provide an unbiased evaluation of the AI technologies employed, helping to identify shortcomings and mitigate potential biases or adverse effects. By embracing such measures, businesses can gain valuable insights and take proactive steps to address any concerns and enhance the overall performance and ethical standards of their AI applications.

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

A. Yes. There are consumer-centric disciplines proposed by research organisations aimed at regulating the appropriate use of AI technologies. To build reliable and safe AI systems, it is crucial to consider proper internal functioning and mature, robust schemes to respond to external threats by third parties. The design of an AI system should address, mitigate and make transparent the shortcuts and biases that stakeholders may be unaware of, aiming to minimise intentional and unintentional algorithmic and social biases, as well as self-interest biases. In this way, transparency plays a vital role in alleviating concerns and fears related to AI systems. Once the technology is adopted, it allows users to ask questions about the AI systems can encourage a positive, realistic, and ethical adoption of their collected data.

AI systems should incorporate a well-designed control panel that provides an appropriate level of autonomy to human users. Additionally, a user-friendly and easily accessible visual or tangible interface should be implemented to enhance usability and accessibility. This could be supplemented by increasing educational awareness of AI technologies through the use of workshops and online tutorials. This would empower users to interact with AI effectively.

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

A. The context-specific nature of the framework and the principles-based approach will somewhat limit these problems. The framework's iterative approach will also be beneficial in this regard. Special attention should be given to the alignment of the UK regulatory framework with similar international efforts, particularly in the EU and the US. As UK businesses need to collaborate with international partners or maintain activity in international markets, a lack of regulatory alignment can lead to contradictory guidance and constrain innovation.

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

A: Strongly agree.

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

A. The best way to measure the impact of the framework is through quantitative and evidence-based risk assessment based on "real-world" data such as historical incident reports and use cases and stakeholders' feedback and surveys. Findings should be communicated clearly to relevant stakeholders to inform future policy.

Q.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?

A. Somewhat agree.

Q.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?

A. Yes.

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

A. Regulators require extra expertise and resources to comprehend the technical and ethical aspects of AI and conduct effective auditing and enforcement activities. Additionally, access to training data is necessary to ensure companies are developing AI fairly and ethically.

Q.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?

A. Somewhat agree.

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

A. Utilising a Corporate Digital Responsibility (CDR) framework towards responding to risk and building public trust by organisations taking actions such as the establishment and adherence to a Digital Responsibility Code, defining and aligning organisational purpose with CDR goals and that of the pro-innovation five principles, such as principle 5 (accountability and governance), through actions such as implementing strong digital governance within organisations in the forms of a Digital Ethics Board, internal reporting and monitoring systems, and CDR champions on each team/department. This guarantees continuous human oversight facilitating principle 2 (appropriate transparency and explainability).

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

A. Whilst the overall approach does a good job of explaining "what" needs to be done under the five principles we recommend organisations engage with the Corporate Digital Responsibility framework as a means to incorporate the five principles long-term. This will allow the creation of an organisational culture that encourages ethical practices around AI in ways that are socially, economically, and environmentally responsible.

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?

A. The key challenge is creating a comprehensive legal framework that can be responsive to technological change and protect users without stifling innovation. Rapid progress in the field of AI keeps producing outcomes and capabilities that were unthinkable a few years ago. This means that the proposed framework must adapt quickly to advances in the field of AI without hindering innovation. Therefore, it is vital that regulators establish a clear feedback loop with industry to ensure the principles are relevant, clearly understood and embedded at the design stage.

L2. 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?

A. Somewhat agree.

How could it be improved, if at all?

A. It could be improved by placing more emphasis on developing a responsible innovation framework to ensure that principles such as security and fairness are considered at the design stage. This should include a mandatory ex ante impact assessment of the new technology to understand potential complex harms that may arise.

L3. If you work for 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?

A. To create an effective AI strategy, organisations should tailor it to their specific needs. This includes setting clear goals, allocating resources, managing risks, and establishing ethical guidelines. It is important to develop a roadmap for implementation, educate and train employees, collaborate with other institutions, stay informed about AI developments, and monitor performance. By doing so, organisations can optimise AI adoption, identify emerging risks, and ensure responsible and successful AI implementation.

The government could establish clear, coherent and coordinated regulatory frameworks addressing AI-related risks and providing guidelines for responsible AI development and deployment. It could institute cross-cutting standards, certifications and requirements for AI systems to which businesses would have to adhere. It could also allocate funding for research and development in AI risk management as well as facilitate collaboration and knowledge exchange within industry.

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?

A. Efficient and high-performing foundation models require significant investments of resources and extensive training efforts. Companies that develop and release these models dedicate substantial resources to ensure their effectiveness while actively implementing safeguards to address potential legal issues and negative publicity. Rather than directly releasing the actual model, they offer a controlled means for users to utilise it. However, if such models are released or leaked, they can be easily fine-tuned by any interested party using various datasets, thereby eliminating the previously implemented safeguards. This grants extraordinary power to individuals with malicious intentions, enabling them to create efficient models for spreading misinformation, operating malicious chatbots, engaging in scams, and other detrimental activities. Consequently, this may lead to the generation of biased, false, or harmful content. Furthermore, complications regarding intellectual property rights pertaining to these models' training data and outputs may also arise.

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

A. Somewhat disagree.

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

A. Due to their transformative nature, foundational model systems such as GPT-4 will span the jurisdiction of nearly every sector regulator. Therefore, to avoid excessive regulatory burden on AI companies and to provide consumers with confidence to use this technology, a cross-sectoral approach to regulation may be more appropriate. This could be accompanied by mandatory licensing and testing requirements, as well as monitoring throughout the AI's lifecycle.

S1. To what extent would the sandbox models described in section 3.3.4 support innovation?

A. Single sector, single regulator: somewhat support innovation.
Multiple industry sectors, single regulator: somewhat support innovation.
Single sector, multiple regulator: strongly support innovation.
Multiple sectors, multiple regulators: strongly support innovation.

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

A. There should be a commitment to periodically evaluating and improving sandboxes based on feedback from AI innovators, regulators, and other stakeholders. In addition, the government must implement mechanisms for providing timely and transparent feedback to participants.

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

A. Participation could be increased by fostering collaboration with Human–Computer Interaction Design (HCI) researchers to develop a comprehensive set of engaging activities that facilitate the active involvement of diverse stakeholders, including users, in the design and evaluation of regulatory sandboxes to better understand the risks and the impact of AI use. Implementing clear guidelines regarding participation and intellectual property rights is also essential to ensure stakeholders feel confident in engaging with the sandboxes.

S4. Which of the following industry sectors do you believe would most benefit from an AI sandbox?

A. Financial services and insurance; Communications; Healthcare; Research and Development.

For inquiries, contact us.

agencyresearch.net
agencyresearch@newcastle.ac.uk