Response to the Department for Science, Innovation and Technology consultation on the white paper 'A pro-active approach to AI regulation' published in March 2023

Authors: Citizens Discussion Group on Social Impacts of AI

contact email: rod.rivers@mac.com

Introduction

Please accept and consider this response. It has emerged from a small group of citizens who have had monthly discussions since November 2022 to consider the social impacts of AI.

Rather than respond to the individual consultation questions we have put our views forward in this free format. This is so that we are not constrained by the assumptions made in the white paper and so that we can respond at a higher level than the detail of the proposals.

Main Points

Firstly. we are struck by the coincidence in time between the consultation white paper and the proposal by AI experts that there should be a moratorium on AI developments. There is clearly an anomaly between these two positions that needs to be fully understood and reconciled. On the one hand we have the UK Government proposing very light-touch regulation in order not to stifle innovation and on the other hand a significant number of AI experts who are alarmed by the existential risk presented by AI systems. The way in which these two rather opposing positions can be reconciled is explored below.

Secondly and related, AI is developing rapidly, and indeed, much has happened in the three months since the publication of the white paper, which itself may have been drafted before the public were fully exposed to systems like Chat GTP. For example, since then Geoff Hinton (the 'grandfather of machine learning techniques') has expressed his concerns about AI and the uncertain potential for AI systems to develop a different but more effective form of intelligence, in which many interacting AIs communicate at electronic speed with access to the accumulation of human knowledge (both well validated and spurious). We are already seeing AIs that, at present with human help, generate computer code that can feed into the development of another, more powerful and effective, generation of AI systems. We can anticipate that it will not be too long before AI systems are critiquing and improving themselves in a general (not task specific) way without much, if any, human intervention. Geoff Hinton, himself, believes that it is very difficult to predict how AI will develop beyond a 5 year horizon. While the white paper acknowledges the pace of development, it offers no mitigation of a risk that might emerge from the fog within the lifetime of a single

government. We cannot see how either a moratorium on developments of AI or its regulation is likely to prevent an AI 'arms race' amongst nations or between companies. This situation calls for a quite different strategy and our proposals below address this.

Thirdly, we note that the UK cannot act alone with respect to development of AI or its regulation. While the UK may aspire to being amongst the leading innovators in AI, it is likely to be dwarfed by the US, China, and the EU, both in its investment in AI and in its regulation. With respect to development this means a potential 'arms race' that can only accelerate developments and heighten the existential risks. With respect to regulation this means that the UK, while it might be able to influence the debate about regulation, could in the end be subsumed into a regulatory regime largely defined by other bigger players.

Fourthly, we note that on first reading the white paper comes across as a 'do nothing' approach (light touch regulation and minimal organisational change). It creates the impression of lacking leadership, firm direction, and moral authority. AI is an area of increasing expert and public concern. It is dominated by multi-national commercial players who have already demonstrated an inability to adopt adequate safeguards in relation to harms caused by technology. Only a state can take the moral lead necessary to control large, short-term commercial interests that often have little interest in minimising harms.

Fifth, the white paper fails to distinguish between the types of AI system that do harm and those that clearly confer benefit. Its laisser-faire approach fails to acknowledge the full impact of either past or future harms. The white paper refers to regulating 'use' rather than 'technology'. We would argue for regulating with respect to 'harms'. There is insufficient explanation as to the purpose of regulation as a mechanism for containing the risk of harms.

We support the principles generally and especially the principle of contestability and redress. Compensation and redress with respect to 'harms' gets directly to the purpose of regulation and would encourage the identification, anticipation and mitigation of harms. Making contestability easy is essential to the fuller understanding of fast changing impacts of technology. Making redress in cases of transgression fully compensate individuals and society (i.e. to compensate for wider social harms) is the only language big players are likely to understand. The burden of proof of no harm (or benefit clearly and significantly outweighing harms) should be on the companies

developing AI and their products rather than on the individual or society. Contestability and redress are far less abstract than principles like fairness and are the route to operationalising what the more abstract principles mean in practice. If legislation is to have any teeth it needs to do everything it can to facilitate and fully compensate for both individual and social harms.

Sixth, whilst we welcome the white paper and the opportunity it offers to address an important topic, we believe that by not adequately taking into account the above, the white paper is flawed because it addresses the wrong question. This is highlighted by the huge gulf in the position of the white paper for light-touch regulation and the perception by experts in the industry that AI may constitute the world's biggest existential threat. There is clearly a significant mis-match in the frames of reference behind these two apparently opposing positions.

We believe that rather than addressing the extent of regulation - whether it should be light-touch or not, it should be addressing the question 'what sorts of AI should the UK be developing?'. In the answering of this question, we point to an important role for the UK for which there may be a gap in the market, and a vital role in the development of AI systems generally.

Proposed revision to emphasis of the white paper

We have a suggestion that we believe should be considered at the highest levels of government, that could steer the UK onto a path that would not only be good for the UK but could also benefit all nations. It avoids the UK becoming implicitly complicit in a dangerous AI arms race, it capitalises on strengths that are deeply routed in British culture, it provides an excellent opportunity for commercial exploitation of AI technology, and it enhances the brand and credibility of the UK as a leader in a matter of importance across the world.

So, what is the suggestion? In one sentence - the UK Government should set up mechanisms that support AI developments designed to prevent harms (and to capitalise on some strengths in ethical big data) and discourage others.

Mechanisms: The mechanisms we propose are financial incentives (e.g. grants and tax-breaks in the case of favoured developments, and

additional tax burdens and regulation in the case of discouraged developments). Such an approach provides a more focused, useful and effective regime than that proposed in the white paper. It operationally defines what is encouraged and discouraged without introducing heavy regulation that will quickly date or stifle innovation. It clearly indicates the areas where the UK can lead the world, and where other nations may have neither the capability or motivation to compete while still seeing the benefits of and supporting the UK's initiative.

Areas to Encourage: While other nations focus on the mainstream development of AI systems, our suggested approach is to encourage focused specialisation on particular AI technologies and applications. These are ethical AI, AI risk mitigation and big data applications (e.g. health).

By coordinating and integrating the skills that UK universities have built up, particularly in the social science, computing and creative industries, the UK can build world-class interdisciplinary teams designed to address some of the hard problems of AI. These include:

- AI safely and checking of AI decision-making
- truth verification, anomaly detection and fraud/fake detection
- accountability, responsibility and legal liability in relation to AI systems
- identity verification, protection and management
- collective intelligence and citizen participation in political decision-making
- legal judgement and moral reasoning
- traceability

Included in the above are : watermarking technologies, reflective and corrective algorithms, architectures to facilitate citizen participation in decision-making, anomaly detection, personal/user agents, cyber defence and neutralisation systems, open decision-making and transparency, blame logics (to help formalise the allocation of accountability, responsibility and redress for harms), explanation systems, scientific truth verification, supply chain traceability (e.g. food, energy clothing)., health and safety checking systems, checking for adherence to law and regulation, and many other specific applications that would help ensure the integrity of AI and other systems.

Rationale and Discussion

The approach has parallels with the way in which the UK has encouraged the development of green / carbon neutral technologies. The impacts and benefits accrue not just to the UK but to individuals everywhere and humanity as a whole. Hence the products and services are welcomed by governments, new businesses and citizens alike. The market for these AI developments is world-wide and wide open.

The UK also has the opportunity to capitalise on the use of high quality big data sets. In particular, data held by the NHS needs to be both protected and exploited by AI. The UK can use big datasets like health data without either compromising privacy or selling the crown jewels. How? These big data sets can be used to train AI systems. The algorithms produced by this training have large commercial value. They do not expose the raw data itself so patient privacy is maintained and the ownership of the raw data is preserved.

The UK has an opportunity to act on the world stage in a stateman-like way with respect to AI. It could have with its eye firmly on helping mitigate the risks of AI systems by building principles, methods, skills and tools that draw on its already strong capabilities in providing legal frameworks, regulatory systems, democratic government and higher education. AI is not just another technology. It is a technology that will have a profound effect on humanity and the UK should position itself as a forward thinker and actor. It need not join a knee-jerk competitive scrabble to develop short-term commercial AI systems – a race that it will surely lose to bigger international players.

Instead, it could develop a critical mass of capability that steers AI towards beneficial applications and provides the tools to mitigate AI risks. The risks have already materialised through the use of AI in marketing, manipulating news feeds, deep fakes, dissemination of false and mis-leading information and so on. We will already face the possibility of the use of AI by organised crime, rogue states and borderline industries like gambling. Look at the way the UK in particular has fallen behind in fraud detection. We need a concerted focus and resources for being ahead of the threats made possible by AI to address both short-term threats like fraud but also the longer term) emerging existential threats (and in this case long term may be only 5 years. The UK could develop a commercial edge by leading in ethical AI, AI risk mitigation and use of big data in training algorithms.

For each of the principles set out in the white paper (safety, transparency, explainability, fairness, accountability, contestability and redress) it is possible to identify/develop scenarios illustrating risk. For example, how might a well-resourced actor build AI systems that threatened safety, transparency etc., either malevolently or inadvertently. Such scenarios might drive the development of AI tools that can 'police' other AI systems and help counter the risks to these worthwhile principles.

All the example areas above require a multi-disciplinary approach. They need to be addressed from both a social and technical perspective. They need to be developed rapidly to address the many threats and risks posed by the ready availability of AI platforms. As we have seen by the use of AI in social media, these new AI platforms can be used in many nefarious ways ranging from the criminal to unacceptable exploitation (both commercial and political). There are also many possible applications of AI that can help address currently intractable UK and world problems. These too can be encouraged.

Recommended changes to the white paper

In order to implement the above we make the following recommendations:

The white paper, while useful as it stands, should be re-oriented (and supplemented) to:

- position the UK as the leading developer of products and services to support the development (and commercial exploitation) of ethical AI, AI risks mitigation and AI training data
- identify ethics as a primary driver of the policy along-side innovation and commercial exploitation
- explicitly set out the areas of AI development where harms and potential harms have already been identified and cite the types of developments and applications that potentially lead to harms
- Propose the development of mechanisms to identify, measure and allocate the accountability / responsibility for harms as a basis for determining appropriate redress
- explicit identify areas of AI development to be encouraged and only use examples that conform to the areas encouraged

- commission the development of a test (that could potentially be implemented as an AI system through training on examples) that would score proposed development for their conformity with the types of development to be encouraged (i.e. ethical and commercially promising)
- set out mechanisms by which these areas might be encouraged (e.g. grants; tax-breaks; technical, training and managerial support)
- set out a strategy to develop a world leading workforce able to develop products and services in the areas to be encouraged
- Develop scenarios that illustrate risks to the principles set out in the white paper (safety, transparency, explainability, fairness, accountability, contestability and redress) and use these to drive the development of AI systems to mitigate the risks
- set out deterrents to discourage potentially criminal, harmful and otherwise unacceptable developments (e.g. heavy regulation and punitive taxation)

Additional Comments

Otherwise, we like the iterative and sandbox approaches, and especially in light of the above comments we support centralised coordination and oversight.

We felt that case study 2.1 relating to the use of AI to determine insurance premiums might send the wrong messages about useful applications. It is arguable that such an application may go against the principle of insurance as a mechanism of fairly spreading risk. In general, the applications used to illustrate the types of AI development that are to be encouraged need to be thought through more explicitly and illustrate how the principles play through into their selection. The case studies should clearly and explicitly exemplify the operation of the principles. Indeed, the AI applications that should be encouraged should be those that would implement the principles – building AIs that aim to achieve greater safety, transparency, explainability, fairness, accountability, contestability and facilitate determining redress are exactly the sort of applications where we should position the UK to become world-leaders. Applications like these would have great value to society, would be commercially valuable in a world where business would benefit from the greater stability they might engender and put the UK in the forefront of AI safety.