WRITTEN EVIDENCE SUBMITTED BY UNIVERSITY OF CAMBRIDGE

GAI0094

We're researchers at the University of Cambridge's Leverhulme Centre for the Future of Intelligence and the Centre for the Study of Existential Risk. These are both interdisciplinary research centres at the University, focussed respectively on making the best of the opportunities of AI as it develops over coming decades, and on studying and mitigating global risks.

We have a longstanding interest in this topic. We've led major research on AI governance such as The Malicious Use of Artificial Intelligence and Toward Trustworthy AI. We first submitted evidence to the Science and Technology Committee in 2016, during the Robotics and artificial intelligence inquiry. We have also welcomed the Lords Select Committee on AI to Cambridge in 2017. We have also had extensive discussion with the UK Government, governments of the USA, Singapore and EU; international organisations such as the UN and OECD; and leading technology companies. Dr Ó hÉigeartaigh is part of GPAI, and LCFI is part of the Partnership on AI.

We fully support the aim to keep the UK at the forefront of responsible innovation in AI, as a 'global AI superpower'. Its recent policy paper is an important signpost of the UK's approach to putting in place a framework for the governance of AI. It is vitally important to get this framework right. If the UK gets its framework right, it will to a large extent dissolve the supposed tradeoffs, promoting consumer and business confidence.

Our main suggestions are:

- 1. Regulatory oversight of AI
    - Establish new institutional architecture to oversee coordination, spot duplication and gaps, and support regulators
    - Agree timelines with regulators
    - Clarify relevant regulators
    - Address foundation models, a current gap
    - Address other gaps
- 2. Principles
    - Add human oversight to principles
    - Add "Ensure robust internal processes" as a principle
    - Recognise the limits of principles
- 3. International
    - Be active in international fora

We would be happy to discuss our response further with the Committee. Please find our answers below, and attached.

- Haydn Belfield, University of Cambridge (Academic Project Manager, Centre for the Study of Existential Risk & Associate Fellow, Leverhulme Centre for the Future of Intelligence)
- Seán Ó hÉigeartaigh, University of Cambridge (Acting Director, Centre for the Study of Existential Risk and Principal Researcher, Leverhulme Centre for the Future of Intelligence)
- Shahar Avin, University of Cambridge (Senior Research Associate, Centre for the Study of Existential Risk)
- José Hernández-Orallo, Universitat Politècnica de València (Professor)
- Giulio Corsi, University of Cambridge (Research Associate, Leverhulme Centre for the Future of Intelligence)

Regulatory oversight of AI

Establish new institutional architecture to oversee coordination, spot gaps and support regulators

The policy paper notes a need to consider "whether new institutional architecture is needed to oversee the functioning of the landscape as a whole and anticipate future challenges".

Our view is that this new institutional architecture will be needed for three main reasons: to provide coordination to reduce overlaps and inefficiencies; to spot gaps and anticipate upcoming challenges; and to support regulators with expertise and guidance.

First, some central or coordinating body or architecture would help oversee coordination. This would reduce inefficiencies and overlaps, which otherwise could lead to lack of business clarity and burdens on innovation. As the paper notes, we must "ensure that organisations do not have to navigate multiple sets of guidance from multiple regulators all addressing the same principle".

Second, spotting gaps and anticipating upcoming challenges. It would not be any of the specific regulators responsibility to spot gaps. This could lead to risks falling through the cracks.

Third, it could support regulator expertise.Some central or coordinating body could share best practice and relevant, timely information between regulators. Such a body could ensure that regulators are following the steer from ministers, for example with "government-issued guidance to regulators" and "supplementary or supporting guidance, for example focused on the interpretation of terms". Many regulators do not have access to the skills and expertise required to regulate in way that fully. The body could also provide training and enable access to external expertise by providing "pooled capabilities" and "secondments from industry and academia." It would likely be inefficient for each regulator to provide training or seek to access external expertise itself.

Fourth, it could anticipate upcoming challenges. For example, it could have responsibility for monitoring and assessing AI R&D. Increasing evidence is building about the rapid pace, and surprising nature, of advances in AI and ML. Regulatory action at the point of AI deployment/commercialisation is thus often too late to mitigate novel harms. Insight and oversight into AI R&D processes could contribute to a better picture of emerging challenges, and help formulate anticipatory governance solutions. To make sure this minimises burden on UK AI innovation, while capturing the main sources of risk, there is room to explore clear reporting criteria that aim to include only highly innovative advances, or advances in high-risk domains. In many cases, mutual exploration of the R&D frontier between academia, industry labs and regulators may be sufficient. It may be worth considering if regulatory backing, though investigatory powers and a clear mandate to observe and monitor the R&D frontier, could help build trust in the AI ecosystem as a whole. This is particularly relevant in relation to foundation models, as discussed below.

Agree timelines with regulators

There are several “areas of of high risk that demand an agreed timeline for regulators to interpret the principles into sector or domain specific guidance”

For example, the paper itself notes that some applications of AI can have “significant impact on safety - and while this risk is more apparent in certain sectors such as healthcare or critical infrastructure”.

Healthcare and critical infrastructure are two clear areas. However, there are several other areas which arguably demand an agreed timeline, such as the use of AI:

- In our schools, for example to assess students.
- By the police, for example for attempting predict crime in particular locations.
- In our courts, for example to assist judges.
- At our borders, for example to examine applications for visas or asylum.

Clarify relevant regulators

It was somewhat hard to work out from the policy paper which regulators were relevant - both in general and in which particular application areas. The policy paper discusses this “patchwork” and touches on the ICO, CMA, MHRA, HSE, EHRC, and Ofcom, but then also hints that BoE & FCA might be relevant. This makes it hard to work out if there are any overlaps or gaps.

It would be helpful for the next iteration, or the white paper, to have some sort of list or map of all the regulators that may be relevant. We recommend that this be done for the entire landscape, and for some particular areas (such as critical infrastructure and healthcare). This would help answer several of the questions in the policy paper - to whom does the policy paper

refer itself to? Which “regulators will lead the process of identifying, assessing, prioritising and contextualising the specific risks”? This would be very useful for addressing one of the ‘Next Steps’: consider “whether any gaps exist in the existing regulator coverage that require a more targeted solution.”

However, two major gaps suggest themselves.

Address foundation models, a current gap

The trend in AI development is to train large, costly, wide-capability “foundation models” or ‘general purpose AI systems’ (such as GPT-3 or PaLM) which are then tailored to (or ‘finetuned’ for) particular tasks and application areas. The downstream applications of these models are likely to impact across several sectors at the same time. Flaws in the original foundation model (such as security or safety vulnerabilities, or biases) are often replicated in downstream applications. Correlated vulnerabilities across many downstream applications can pose a systemic risk.

Foundation models would currently likely not be captured by the current framework. For example, it is excellent that the case study explores a startup providing a legal, financial or medical advice chatbot “built on top of a Large Language Model”. Security or safety flaws in the underlying model could be replicated in each of these different application areas. However, it is unclear which regulator, if any, would be relevant to the foundation model itself.

This gap poses near-term real risks. Moreover, it also disadvantages UK innovators. This is because not tackling foundation models puts the burden on SMEs relative to big developers. Foundation models are mostly currently coming from the US and SMEs will be tailoring them for a UK context, so this disproportionately hinders UK businesses compared to US ones.

One type of foundation model are generative AI models that produce AI-generated images, text, or video. All generative AI models have the potential to produce credible, high-quality misleading content at scale and at a low cost, potentially creating problems for online information spaces. These services, such as OpenAI’s DALL-E and GPT-3 have typically been offered through an API, through which the company is able to nudge outputs in a less biased, less toxic, safer direction - and to track usage by potential bad actors. However, so-called ‘open source’ companies have risen in prominence over the last year or so. One example is stability.ai, headquartered in London, which just raised $101 million. The impacts of these industry changes are worth exploring: lowering the barriers to generating AI images and text potentially increases access but may have downsides such as lower moderation and security checks. Publication and release norms in AI are an area of concern that is no longer theoretical. Different labs and groups are picking very different approaches with different risk tolerances, and it would be valuable for government to monitor this space and increasingly build and encourage agreement on appropriate levels of harm reduction.

Address other gaps

The paper helpfully notes potential gaps:

“As current UK legislation has not been developed with AI in mind, there may be current risks that are already inadequately addressed, and future risks associated with widespread use of AI that we need to prepare for. For example, around the need for improved transparency and explainability in relation to decisions made by AI, incentivising developers to prioritise safety and robustness of AI systems, and clarifying actors’ responsibilities. There is also concern that AI will amplify wider systemic and societal risks, for instance AI’s impact on public debate and democracy, with its ability to create synthetic media such as deepfakes."

We endorse this list. For example, we warned about the impact of AI on public debate and democracy back in 2018 in The Malicious Use of Artificial Intelligence, and have subsequently explored it in Epistemic Security.

We would also note that the risks around dual-use scientific research would likely be a gap. A 2022 Nature paper demonstrated that an AI system for drug discovery could also produce potential weapons. It is unclear which regulator’s responsibility this would fall under.

2. Cross-sectoral principles

The cross-sectoral principles are a useful starting point as an indication of priorities and values to the regulators who “will lead the process of identifying, assessing, prioritising and contextualising the specific risks”. Three important points:

Add human oversight to principles

‘Human oversight’ is common in many other sets of principles. This obviously has to be context-dependent - the type of human-in-the-loop in which human approval is needed before every action will not be necessary for every application. Human oversight could perhaps fall under ‘Ensure that AI is used safely’ or ‘Define legal persons’ responsibility for AI governance’ but it seems clearer to break it out as its own principle.

This builds on OECD AI Principle 1.2 b) “implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context”.

Add “Ensure robust internal processes” as a principle

A principle to cover some more ‘administrative’ elements would be useful, such as companies having a good risk management system, legible technical documentation, and appropriate record-keeping of the development and deployment process. The principle could be “Ensure robust internal processes”.

This builds on the OECD AI Principles 1.4 b) “ensure traceability, including in relation to datasets, processes and decisions made” and c) “apply a systematic risk management approach”.

Recognise the limits of principles

While a useful starting point and indication of priorities and values, principles have limits. Over 80 sets of principles have been published worldwide (often with very similar content) yet few of these engage with how these principles may be in tension with one another. This highlights the importance of a new institutional architecture (see below) that can provide coordination to reduce overlaps and inefficiencies, spot gaps and anticipate future challenges, and support regulators with expertise and guidance.

3. International

Be active in international fora

One of our major concerns is EU-US standard-setting that might not be well-suited to the UK's needs.

The EU AI Act is likely to have a significant *de facto* Brussels Effect. Their obligations will be fairly stringent, and ‘cost of differentiation’ may be quite high, so many tech companies will just make sure they are in compliance with EU standards and not bother training other systems for the European market. (Technically these standards would be ‘common specifications’ established by CEN-CENELEC and approved by the Commission.)

In the US, the NIST AI Risk Management Framework and Playbook is likely to develop technically detailed standards with a lot of buy-in from the Big Tech companies.

There is a clear channel for the EU and US to come to some overall level of agreement: the EU-US Trade and Technology Council. They will also have a strong incentive to try and set global standards through that channel, in order to exclude China. Another channel is *de facto* alignment as the big tech companies will engage closely with NIST and the EU.

However, the concern is that this might exclude the UK from decision-making, and might lead to these standards not being as well-suited as they could be to the UK’s needs (in terms of “interoperability” and “British values” as the paper notes).

Developments such as the AI Standards Hub and CDEI assurance roadmap are to be welcomed. The paper also helpfully emphasises the importance of working through international fora such as CoE, OECD, GPAI, ISO/IEC. These are very important. We will need to be especially active in the fora we are involved in, and seek to enter into or influence the fora we're not in.

(November 2022)