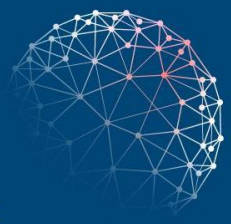

THE CENTRE FOR
LONG-TERM RESILIENCE

Response to "Establishing a pro-innovation approach to regulating AI"

PREPARED BY

Dr Jess Whittlestone

Head of AI Policy. Centre for Long-Term Resilience

THE CENTRE FOR
LONG-TERM RESILIENCE

Executive Summary

We agree with the policy paper that a context-driven approach to AI regulation has many advantages. These include being sensitive to the fact that the risks posed by an AI system are often heavily dependent on context; drawing on the strengths of the UK's existing regulatory ecosystem; and minimising barriers to beneficial innovation.

However, we also believe the proposed approach faces several important challenges, which need to be carefully addressed:

1. **Promoting coherence and reducing inefficiencies across the regulatory regime**. Although regulatory responses are likely to differ by sector, it is also likely that regulators will face many similar challenges where consistency and sharing of information will be important.

2. **Ensuring existing regulators have sufficient expertise and capacity**. Understanding the regulatory implications of AI in even a single sector is not straightforward, and regulators will need training, the ability to share best practice, and access to external expertise to do this well.

3. **Ensuring that regulatory gaps can be identified and addressed**. There will inevitably be important areas of harm from AI that do not neatly fit within the remit of existing regulatory bodies, such as possible applications of AI to dual-use scientific research.

4. **Being sufficiently adaptive to advances in AI capabilities**. We are seeing rapid progress in AI capabilities, and a trend towards increasingly general-purpose systems. The regulatory regime must be able to keep up to date with advances and their implications for regulation.

Cross-sector principles will be a useful starting point for guiding the regulatory regime, and especially for promoting coherence, but much more will be needed. In order to successfully navigate the above challenges, we believe it is crucial that the regulatory regime is supported by a broader governance ecosystem which can effectively identify and address inefficiencies and gaps. In practice, this means **identifying actors with a clear mandate and the capacity to address challenges 1-4 outlined above**. We do not necessarily think this needs to be a single body, since different actors may be

well-positioned to fill different gaps, and an ecosystem of bodies with different responsibilities and powers may be most effective in practice.

In our view, one of the biggest challenges for this regulation will be **keeping pace with the speed of AI progress.** To address this challenge, we recommend that:

- A central government body such as the CDEI or Office for AI should **invest in infrastructure to monitor AI inputs and progress**.
- The Office for AI should explore how **foresight and anticipatory governance methods can support the regulatory approach**.
- The whitepaper should consider **specific challenges that ongoing progress in foundation models may pose** for a context-specific approach to regulation.

Below we provide a more detailed response to questions 1-4 from the policy paper.

1. What are the most important challenges with our existing approach to regulating AI? Do you have views on the most important gaps, overlaps, or contradictions?

The challenge of keeping pace with AI progress

One of the biggest challenges for this regulation is how to keep pace with the speed of developments in AI. **A particularly important challenge is posed by recent progress in foundation models**: models trained on broad data that can be adapted to a wide range of downstream tasks (Bommasani et al 2021). It is likely that in coming years, an increasing proportion of societal applications of AI will come from downstream applications of foundation models, which could challenge the UK's existing approach to AI regulation in a few ways:

- **Regulating sector-specific applications may become increasingly inefficient**, as progress in foundation models leads to very rapid diffusion across a range of sectors at once, and as potential harms may stem from underlying

characteristics of the foundation models (e.g. bias or limitations of datasets) which are shared across many contexts;
- **It will become increasingly important to distribute responsibility across the entire supply chain of AI development.** Only regulating at the application level may hinder innovation by putting an increasing and undue burden on SMEs relative to developers of foundation models, as those providers often will not have the capacity, tools, or access to information about the training of foundation models needed to meet requirements.
- **Foundation models may also precipitate unanticipated harms on a broader societal level** which sector-specific regulators are not well-placed to respond to. For example, safety and security flaws in the underlying foundation model can be replicated in models based on (or finetuned from) them, thus creating a systemic vulnerability.

The UK's approach to AI regulation must acknowledge and begin considering how to address these challenges from increasingly general-purpose AI systems such as foundation models. In particular, we suggest that the UK's regulatory approach should:

- **Recognise that some form of regulation may be needed for general-purpose systems such as foundation models in future**, and commit to exploring the details and feasibility of such regulation further.
    - There are many ways that developers of foundation models (or other actors further up the supply chain) might be held appropriately accountable for harms from AI development without hindering innovation. For example, we might require developers of foundation models to **cooperate with regulators to identify and prevent cases of potential misuse**. We might require developers of foundation models to **follow and document certain best practices** to ensure that systems function reliably and robustly and make use of sufficiently representative datasets. This could reduce the regulatory burden on smaller businesses: a business using a foundation model which has already met certain requirements might then be subject to fewer requirements on the application level.
- **Consider how other policy levers beyond regulation could be used to better understand and manage the challenges posed by foundation models** and other general-purpose systems. We recommend that the government:
    - **Improve monitoring of progress in foundation models and other powerful AI systems.** This might include collecting data on trends of

compute usage (which is likely to be indicative of where powerful general-purpose systems are being developed) and keeping a registry of foundation models and any known incidents or issues with these models which could affect downstream applications.
    - **Support sharing and development of best practices** for the safe and responsible development of foundation models and other general-purpose AI systems (possibly requiring clarification and guidance from the CMA in specific areas, similar to current proposals on how the competition and consumer regimes could better support the UK's net zero and environmental sustainability goals). These could be developed initially into voluntary guidance and later codified into standards.
    - **Contribute to discussions in international fora** around the need for standards for safe and responsible development of foundation models, and advocate for international standards-setting processes around such models.
    - **Incentivise research to develop tools for auditing** or interrogating foundation models or other types of general-purpose AI systems, for example via competitions.

Beyond the specific challenges posed by foundation models, it will be important for regulators to understand and anticipate how trends in AI development may introduce the need for new forms of regulation or challenge existing approaches. To support the regulatory approach, we therefore recommend that the government should invest in:

- **Mapping and monitoring of important AI inputs and outputs.** A central government body such as the CDEI or Office for AI should collect the best information possible on:
    - Important inputs into AI progress, including where large amounts of compute are being used to train AI systems, large or potentially high risk datasets, and algorithmic progress in key areas.
    - Where AI models are being deployed in contexts affecting UK citizens and businesses - for example by establishing a model registry for applications in high-risk sectors.

  Collecting this information would make it easier for the government to identify when and where AI is likely to have particularly large societal impacts; helping to focus regulatory efforts where they are most needed while also making it easier to intervene earlier and in lighter touch ways where possible (Whittlestone and Clark 2021).

- **Scenario development, forecasting, and foresight work.** While we cannot predict the future of AI, many tools and methods exist to map out different possible futures, threats, opportunities, and intervention points. This work will be crucial to ensuring that any regulatory approach to AI is robust to uncertainty about the future, and avoids unexpected surprises (Avin 2019). We recommend that the Office for AI should work with bodies like GO-Science already doing technology foresight work, and organisations like Nesta who are pioneering participatory approaches to anticipatory regulation (Armstrong and Rae 2017; Armstrong, Gorst, and Rae 2019), to **explore how foresight methods and anticipatory governance approaches can most effectively support effective AI regulation**.

Potential gaps in the current approach

As the policy paper acknowledges, "there may be current risks that are inadequately addressed, and future risks associated with the widespread use of AI that we need to prepare for." We agree that **wider systemic and societal risks, such as the impact of AI on public debate and democracy, are particularly important and may not be sufficiently well-covered by existing regulatory bodies** (Seger et al. 2020).

Other potential gaps include:

- **The application of AI to dual-use scientific research**, where it could enable or speed up the development of dangerous technologies (Clarke and Whittlestone 2022). A recent Nature paper showed how AI technologies for drug discovery could be misused to design new biochemical weapons (Urbina et al. 2022). To mitigate the risk of this kind of misuse, regulation may be needed around certain applications of AI to scientific research with dual-use potential, but it is not clear where responsibility would fall for this within the proposed regulatory approach.
- **Threats to critical national infrastructure and cybersecurity** (such as those discussed in Brundage, Avin et. al. 2018) : though the policy paper acknowledges such risks as important to mitigate, it is not clear that there is an existing regulatory body that is well-placed to address such risks, especially since doing so may require considering how vulnerabilities may emerge from interconnected systems.

We recommend that the forthcoming whitepaper provide more clarity on where responsibility for addressing these harms, and broader systemic and societal harms, will lie. It would also be useful for the whitepaper to include a mapping of which regulators are expected to be relevant to AI regulation (and ideally, which regulators will be responsible for specific areas of harm), which would make it easier to identify where gaps or overlaps are likely to arise.

Defining the scope of the regulatory approach

We are generally supportive of defining the scope of the AI regulatory framework in terms of core characteristics, to strike a balance between the need for flexibility in the definition, and the need to put some boundaries on the scope of the regulation.

However, we find some aspects of the characteristics laid out, and how they would be used in practice to define the scope of the regulation, a little confusing.

As stated, the first characteristic ("adaptiveness") seems to very closely match the definition of modern machine learning systems: systems that are not programmed according to explicit rules but rather learned from data, meaning their behaviour often cannot be easily or reliably predicted or explained. Given this, it is not clear that there is any benefit to trying to point to an underlying characteristic here, and it seems simpler and clearer in practice to simply state that any system using machine learning is within the scope of the regulation. If there is a desire to exclude certain types of ML from the regulation, these exceptions could be added. Similarly, if there is a desire for flexibility, so that potential future technologies with a similar level of adaptiveness to modern ML systems would be captured, then this could also be added as a clause (but again, given the broad definition of machine learning and how closely it matches the current description of adaptiveness, it is not clear what such systems would be.)

It also isn't clear whether a system would need to meet both characteristics outlined (adaptiveness and autonomy) in order to fall within the scope of the regulatory approach. It seems that in many cases, the first characteristic - i.e. the challenges associated with modern machine learning systems - is sufficient to pose a regulatory challenge even if there is little to no autonomy in the system. For example, a machine learning algorithm used to make predictions informing medical care should probably be subject to various bias and robustness checks even if it never autonomously makes decisions, because faulty predictions are still likely to inform doctor's decisions in ways that could cause harm.

2. Do you agree with the context-driven approach delivered through the UK's established regulators set out in this paper? What do you see as the benefits of this approach? What are the disadvantages?

We see several advantages of the context-driven approach, as laid out in the policy paper: it allows for the fact that the risks of an AI system often depend heavily on the context of application; it draws on the strengths of the UK's existing regulatory ecosystem; and creates minimal risk of unduly hindering innovation. Another benefit we see, as compared with the EU's horizontal regulatory approach, is that it seems likely to result in clearer, more concrete guidance for businesses operating in specific sectors, and therefore greater regulatory certainty.

However, this approach also faces many challenges, many of which are also acknowledged in the policy paper:

- **Ensuring coherence** across the whole regulatory regime will be a challenge for the sector-specific approach, and inefficiencies are likely to arise where there are similarities between the regulatory challenges faced in different sectors.
- Existing regulators may be constrained in their ability to address the challenges posed by AI by both **lack of expertise and capacity,** a challenge which is likely to become more acute as time goes on and applications of AI become more widespread.
- Individual regulators may be **more vulnerable to being captured by industry** interests than a cross-sector regulator with more wide-ranging powers might be, especially since industry interests in a single sector are more likely to pull in one direction.
- It is likely that **regulatory gaps will arise** where harms from AI do not clearly fall within the remit of any existing regulator or cut across several, and without any actor responsible for identifying and addressing such gaps important issues may be missed.

- Finally, it may be particularly difficult for this context-driven approach to be **sufficiently anticipatory** given the pace of development in AI progress, as discussed in more detail in our response to question 1.

We believe that it is possible to overcome these challenges, but only if the context-sensitive regulatory approach is supported and informed by a **broader governance ecosystem with adequate powers and responsibilities to address capacity and expertise shortages, identify regulatory gaps, and anticipate future issues**. We outline what this broader ecosystem might need to look like in practice in response to question 4.

3. Do you agree that we should establish a set of cross-sectoral principles to guide our overall approach? Do the proposed cross-sectoral principles cover the common issues and risks posed by AI technologies? What, if anything, is missing?

We agree that cross-sector principles will be useful for guiding the overall regulatory approach: principles are often a useful starting point for more precise rules and requirements, and can also underpin important cultural norms and values (Seger 2022).

However, **we would emphasise that principles are only a starting point**, and much more will be needed to ensure that regulators have the guidance they need to do their jobs in an effective and coherent way.

First, practical guidance around the implementation of the principles will be crucial. The same principle can be open to many different interpretations - there are many different dimensions of what it means for a system to be 'explainable', for example, and which are most important may vary by context (Nyrup and Robinson 2022). It will therefore be important to provide some guidance about how to interpret a given principle in different sectors. Similarly, principles may come into tension with one another in practice, or with other goals (Whittlestone et al. 2019) - regulators will have to balance the requirements of safety and fairness against the broad goal of promoting innovation, for example, which may not be straightforward.

Second, other forms of guidance for regulators may be at least as important and useful as ethical principles. For example, it may be useful to provide regulators with a framework or guidance for thinking about the kinds of harms that can arise from the use of AI systems, and new types of harm that may arise from near future

developments in AI. This type of guidance may have some overlap with ethical principles (e.g. highlighting potential harms from bias) but focus on a slightly lower level of analysis which makes it easier for regulators to identify issues they need to address.

4. Do you have any early views on how we best implement our approach? In your view, what are some of the key practical considerations? What will the regulatory system need to deliver on our approach? How can we best streamline and coordinate guidance on AI from regulators?

As outlined in response to question 2, we think one of the most important aspects of successfully implementing the regulatory approach will be **ensuring that a broader governance ecosystem is able to identify and address inefficiencies and gaps**. Cross-sector principles are an important starting point for ensuring coherence across the regulatory regime, but much more than this will be needed.

In practice, this means identifying a body (or bodies) with a clear mandate and the capacity to:

1. Provide **AI-related expertise** and upskilling for regulators
2. Identify cross-sector lessons and promote **coherence**
3. Identify and anticipate **regulatory gaps** and ensure they are acted upon

The "Common Regulatory Capacity for AI" report, published alongside the policy paper by the Alan Turing Institute, discusses how to fill these gaps. Though we think the report raises many useful considerations, **we think it is a mistake to assume that a single body should necessarily take on all of the above responsibilities**. For example, though an independent body like the Alan Turing Institute may be well-suited to acting as a central hub of AI expertise for AI regulators, identifying and acting on

regulatory gaps seems better positioned within a central government body like the CDEI with a closer relationship to regulators.[1]

We think that "common regulatory capacity for AI" needs to look like a broader governance ecosystem, composed of several different actors with different responsibilities and access to different policy levers, rather than a single, independent body which may not have the power to act on its findings or support regulators in the way that they need.

The "three lines of defence" (3LoD) model, which is considered best practice in industry risk management, may also be helpful to consider here. Key to the 3LoD model is the idea that responsibility for risk management should be assigned on three levels: (1) a *first line* which is responsible for day-to-day risk management in specific areas; (2) a *second line* which provides expertise and support to the first line, while monitoring for challenges; and (3) a *third line* which provides independent scrutiny to ensure the first two lines are functioning as intended. Ensuring these three lines are covered could be extremely useful for ensuring that the AI regulatory regime is functioning as intended, with sector-specific regulators playing the first line function, some combination of other bodies such as the CDEI, Office for AI, and Alan Turing Institute serving as the second line, and some form of external audit of the regulatory process as the third line.

We recommend that the Office for AI **commission further independent analysis of how responsibilities 1-3 can best be addressed** within a broader government ecosystem. We also recommend that the forthcoming whitepaper **clearly delineate roles and responsibilities** for this broader ecosystem.

ACKNOWLEDGEMENTS

This response benefited from conversations and input from Markus Anderljung (Centre for the Governance of AI), Dr Shahar Avin (Centre for the Study of Existential Risk), Haydn Belfield (Centre for the Study of Existential Risk), Dr Rune Nyrup (Leverhulme

[1] We also have concerns about the Alan Turing Institute's independence in answering questions about how to address common regulatory capacity - it is striking that the report does not acknowledge the possibility of interviewer bias when reporting interviewee's opinions about the kind of body best placed to provide certain functions, and opinions about the ATI in particular.

Centre for the Future of Intelligence), and Charlotte Stix (Eindhoven University of Technology)

REFERENCES

Armstrong, H., & Rae, J. (2017). A working model for anticipatory regulation. London: Nesta. Available from: https://media. nesta. org. uk/documents/working_model_ for_anticipatory_regulation_0_TpDHt7z. pdf.

Armstrong, H., Gorst, C., & Rae, J. (2019). Renewing regulation.

Avin, S. (2019). Exploring artificial intelligence futures. Journal of AI Humanities, 2, 171-193

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

Clarke, S., & Whittlestone, J. (2022, July). A Survey of the Potential Long-term Impacts of AI: How AI Could Lead to Long-term Changes in Science, Cooperation, Power, Epistemics and Values. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 192-202).

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: a context-sensitive framework for explainable medical AI. Ethics and Information Technology, 24(1), 1-15.

Seger, E., Avin, S., Pearson, G., Briers, M., Heigeartaigh, S. Ó., Bacon, H., ... & Weller, A. (2020). Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world.

Seger, E. (2022). In Defence of Principlism in AI Ethics and Governance. Philosophy & Technology, 35(2), 1-7.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence, 4(3), 189-191.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019, January). The role and limits of principles in AI ethics: towards a focus on tensions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 195-200).

Whittlestone, J., & Clark, J. (2021). Why and How Governments Should Monitor AI Development. arXiv preprint arXiv:2108.12427.