Submission of Feedback to the European Commission’s Proposal for a Regulation laying down harmonised rules on artificial intelligence

Sam Clarke,[1] **Jess Whittlestone,**[1,2] **Matthijs Maas,** [1,2] **Haydn Belfield,** [1,2] **José Hernández-Orallo,**[1,3] **Seán Ó hÉigeartaigh**[1,2]

[1] Leverhulme Centre for the Future of Intelligence, University of Cambridge
[2] Centre for the Study of Existential Risk, University of Cambridge
[3] Universitat Politecnica de Valencia

Introduction

We are a group of academic researchers on AI with positions at the University of Cambridge’s Leverhulme Centre for the Future of Intelligence (LCFI) and Centre for the Study of Existential Risk (CSER), and the Universitat Politecnica de Valencia. We have published dozens of academic papers and reports on the ethics and governance of artificial intelligence. Our past work has included submissions on the HLEG’s Draft Ethics Guidelines and the White Paper. We have had many useful discussions with the Commission, including presenting at the 2020 AI Alliance Assembly at a session chaired by Eric Badique.

We are generally very supportive of the proposed regulation, and particularly the risk-based approach, which recognises the varying levels of risk posed by AI systems in different contexts. We are optimistic that, as the first legal framework for AI worldwide, the proposed regulation can help set standards internationally for enabling the benefits and mitigating the risks of this powerful set of technologies. However, the impact this regulation has will depend on many specific details of implementation, and on how it can interact with other evolving parts of the AI governance ecosystem.

As researchers focused on the long-term societal impacts of AI, we are particularly concerned with how and to what extent the proposed regulation can:

- **Adapt to new capabilities and risks as they arise.**
  Historically, regulation has often struggled to adapt as technologies evolve (Crootof, 2019; Picker, 2001). Ensuring adaptive regulation is particularly important - but also particularly challenging - for AI, where technological progress is rapid, interacts with other emerging technologies, and can have wide-ranging impacts.

- **Manage the broader societal impacts of AI.**
  Like all technologies, AI may cause societal-level harms, even if their direct impact on individuals is minimal. For example, the use of AI to generate fake content online may reduce overall trust in scientific information in ways that could cause downstream harms, without harming individuals directly.

In this submission we make some recommendations for how the proposed regulation could better address these two points, both in terms of (a) how the regulation classifies ‘high-risk’ systems, and (b) what requirements the regulation subjects those systems to. However, we also recognise that regulation is not always the most appropriate tool for addressing longer-term societal challenges. We therefore also make some suggestions for where the broader AI governance ecosystem may need to be strengthened to complement the proposed regulation, and how the European Commission can enable and respond to these changes.

In summary, our recommendations are to:

- Empower the Commission to add high-risk AI systems to the list in Annex III that:
    - are under areas outside of those listed in points 1-8
    - pose a substantial risk of societal harm
- Consider how to identify and regulate high-risk uses of general purpose systems
- Consider how to regulate companies that may intentionally underemphasise the role that AI plays in their decision-making
- Authorise the European Artificial Intelligence Board to propose changes to the regulation's Annexes
- Schedule specific timeframes on which to consider revisions to which AI systems are covered by the regulation
- Broaden the current requirements to include evaluation of broader societal harms, beyond risks to health and safety or fundamental rights
- Be aware of the risk of common specifications being 'captured' by industry, and maximise participation in the setting of these specifications to reduce this risk
- Provide even further clarification and indicative examples in the Chapter 2
- Connect the proposed regulation to a broader governance ecosystem, for example by:
    - Creating more channels for filing complaints to the Commission and notifying authorities
    - Ensuring that some Board members have expertise in AI monitoring, assessment, and horizon-scanning
    - Maintaining an 'AI safety incidents' database at a European level
    - EU institutions and Member States funding more work (within government and academia) on AI monitoring and assessment
    - Ensuring that the Commission can update Annex VIII

Table of Contents

***Introduction* 1**

***Recommendations on classification of AI systems as high-risk* 3**

Broaden the provision to add new high-risk systems 3

Ensure the classification stays relevant as uses of AI develop 4

Implement mechanisms to enable adaptability in practice 4

***Recommendations on requirements for high-risk systems* 5**

Expand requirements to include evaluation of broader societal harms 5

Full participation in common specification design processes.. 5

Provide further clarification and examples in Chapter 2... 5

***Connecting the proposed regulation to a broader governance ecosystem* 6**

Recommendations on classification of AI systems as high-risk

As AI capabilities and deployments evolve, new areas of risk will inevitably emerge, and ways of regulating AI may need to change. It is crucial that the proposed regulation can respond to these risks as they arise, in order to avoid becoming overly narrow or outdated.

To its credit, the regulatory proposal already includes a range of mechanisms aimed at ensuring adaptability, including the general risk-focused orientation, broad definition of AI, and the ability to revise and expand the list of both AI techniques (in Annex I) and the list of high-risk AI systems (in Annex III). However, we believe that more can and should be done to ensure the regulation can adapt as the risk profile of AI changes.

We suggest that the Commission:

Broaden the provision to add new high-risk systems

We recommend broadening the provision to add new high-risk systems and risk areas. Specifically, **we suggest modifying Article 7 to empower the Commission to add high-risk AI systems to the list in Annex III:**

- **under areas outside of those listed in points 1 to 8 of Annex III, or**
- **if the AI systems pose a substantial risk of societal harm.**

Currently, the regulation only allows the addition of new high-risk AI *systems* if they both fall under any of the eight listed areas and are deemed to pose at least as great a risk (to health and safety or adverse impact on fundamental rights) as systems already in Annex III. We see two limitations imposed by these conditions. Firstly, while the eight domains listed are broad, we do not believe they exhaust the range of domains within which AI systems may come to have significant impacts on citizens' lives. AI systems' use in various other domains could raise significant additional risks that are not well captured by these eight risk areas. For example, AI-based personal digital assistants could be used to give individuals important financial, legal, or medical advice with significant consequences for health and safety, and do not appear to be covered by the current risk categories. Moreover, the general purpose nature of AI technology and its rapid rate of progress makes it difficult to anticipate impacts in advance. In light of this uncertainty, it makes sense to have provision for adding areas. Therefore, we suggest empowering the Commission to add high-risk AI systems under areas outside of those listed in points 1 to 8 of Annex III. This can easily be done by removing point 1(a) from Article 7.

Secondly, while many risks from AI technology can be thought of as the potential for harms to *an individual's* health and safety or of adverse impact on their fundamental rights, AI may also cause significant harm, on a *societal* level. For example, digital personal assistants could be used to promote certain products, services, or even ideologies well above others, with the potential to contribute to substantial and potentially harmful shifts in our markets, democracies, and information ecosystems. However, the impacts on individual health, safety, or fundamental rights may be negligible or difficult to discern. We believe the proposed regulation should include provision to identify and regulate systems which could pose this kind of harm, and therefore suggest empowering the Commission to add high risk AI systems to Annex III that pose a substantial risk of societal harm. One concrete way to do this could be to broaden the interpretation of 7.2.d ("the potential extent of such harm or such adverse impact, in particular in terms of its intensity and its ability to affect a plurality of persons") to include societal as well as individual harms.

Ensure the classification stays relevant as uses of AI develop

How 'high risk' AI systems are identified and classified for the purpose of regulation may also need to adapt over time. We highlight two things it may be particularly important for the Commission to consider in this regard:

- **The emergence of increasingly general-purpose systems, which are then adapted and deployed for specific domains by secondary developers or users.**
  Increasingly general purpose AI systems, such as OpenAI's GPT-3 or DeepMind's MuZero, are emerging and likely to become increasingly important in the AI industry. These general purpose tools will increasingly be used by secondary providers who tailor an "off the shelf" product to a specific purpose - for example, a provider could fine-tune a general language model for use in a setting that is considered high risk, such as making decisions in a legal or healthcare context. It is unclear here whether the mandatory requirements should be fulfilled by the provider of the general purpose system or the downstream provider, a lack of clarity which could result in loopholes. We strongly recommend that the Commission consider how to identify and regulate high-risk uses of general purpose systems now, before these systems become more dominant in the market.

- **The possibility that companies may underemphasise the role AI plays in their decision-making in order to avoid regulation.** One way providers or users may seek to avoid the obligations placed upon them is by presenting the use of AI systems as marginal in their decisions, when in fact AI systems are very important. For example, a company could use an AI system to generate some insight, destroy the model but allow human decision-makers to apply the knowledge. In such a case, we believe that the provider should be subject to the same requirements, but determining whether and to what extent an AI system has been involved in a decision may be very difficult to determine. These considerations will become more and more important as and if 'hybrid' human-AI systems, or collective intelligence systems, develop and are used more widely.

Implement mechanisms to enable adaptability in practice

We also recommend establishing concrete mechanisms to enable adaptability in practice. Even where provisions for adaptability exist in principle, historical experience suggests that updating regulations frequently or quickly enough can be challenging. For instance, various arms control regimes have struggled to update control lists in a frequent and timely fashion (Nelson, 2019).

Specifically, we suggest that:

- **The European Artificial Intelligence Board be authorised to propose changes to the regulation's Annexes,** including the list of restricted and high-risk systems. It should also be involved in the formation of common specifications and harmonised standards.
- **The European Commission schedule specific timeframes on which to consider revisions** to the Annexes and any other parts of the regulation.

Recommendations on requirements for high-risk systems

The regulatory proposal lays out a number of sensible requirements for systems classified as high-risk, though these are still relatively high level, and there are currently many different approaches to assessing and assuring the operation of AI systems. The proposal is likely to prompt considerable experimentation in the development of assessment processes over the coming years, which we see as a positive step, and we are pleased to see the proposal include the ability to adopt common specifications to demonstrate conformity for high-risk systems. We make some additional recommendations for how the EC can ensure these assessment processes are sufficient to govern the potentially wide-ranging societal impacts of AI.

Expand requirements to include evaluation of broader societal harms

First, **we recommend expanding the current requirements to include evaluation of broader societal harms, beyond risks to health and safety or fundamental rights**. Specifically**,** we suggest adding a step to the risk management system (Article 9) requiring evaluation of broader societal harms from AI systems. We also suggest including possible societal impact in the requirements for technical documentation (Annex IV). As argued above, AI systems can pose broader societal harms as well as harms to individual health and safety or fundamental rights. The inclusion of this requirement would ensure that these risks are duly assessed by the providers of AI systems.

Full participation in common specification design processes

Second, regarding common specifications, **we recommend full participation in processes by the AI Board and by civil society to mitigate any risk of specifications being overly influenced or 'captured' by industry.** Adopting common specifications will make it easier for providers to demonstrate and regulators to check conformity, and will give providers more legal certainty. However, such specifications do risk being captured by the providers of AI systems, either at the design or implementation phase. To give a concrete example: suppose some company is developing a personal digital healthcare advisor, and knows its recommendations are less accurate for a certain demographic. At the design phase, the company might advocate for a fairness metric which obscures the issue (since there are currently no 'gold standard' metrics for AI fairness); at the implementation phase, they could also modify the test dataset until the discrepancy was not obvious. We suggest that the Commission reduce this risk by maximising participation in the setting of common specifications - though involving the Board in this process, and allowing civil society to attend meetings and submit proposals. More participation makes it harder for specifications to be captured at the design phase, and less likely to leave loopholes that can be exploited in the implementation phase.

Provide further clarification and examples in Chapter 2

Third, **we recommend providing even further clarification and indicative examples in Chapter 2** (mandatory requirements). This would help shape and guide the standardisation process properly. For example, in Article 15 (Accuracy, robustness and cybersecurity):

- 15.3 could be further clarified and specified by including the terms used in the technical literature (see here and here). We recommend adding:
  "The robustness of high-risk AI systems may be furthered through measures to ensure specification, assurance and alignment".
- Adding further indicative examples would also be useful. For example, in Article 15.3, ("The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans"), we recommend adding:

”This may include mechanisms that prohibit some unexpected system behaviours, including preventing the system from operating, if inputs or outputs fall outside a predefined “safe” range."

- Article 15.4 currently reads: “The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent and control for attacks trying to manipulate the training dataset (‘data poisoning’), inputs designed to cause the model to make a mistake (‘adversarial examples’), or model flaws.” We explain and justify other concrete ‘mitigating measures’ to attacks and manipulation attempts in Toward Trustworthy AI: Mechanisms for Supporting Verifiable Claims. We recommend adding:
  “These mitigating measures, may also include red-teaming, bias or safety ‘bug bounties’, ensuring hardware security, threat modelling, and testing in an adversarial setting”.

Connecting the proposed regulation to a broader governance ecosystem

For the proposed regulation to be able to adapt to new capabilities and risks as they arise, and manage the broader societal impacts of AI, it is crucial that it be connected to a broader AI governance ecosystem which can:

- **Conduct ongoing monitoring of AI progress and impacts**.
  - For the regulation to adapt, it needs to not just have provisions to do so easily, but more fundamentally needs to be able to react to high-quality and timely information about how AI capabilities, applications, and their potential risks are evolving.
- **Develop new standards, measures, and methods for assessing AI systems**
  - The ability of the regulation to mitigate harms of AI is only as good as our approaches to assuring/assessing AI systems. Current methods are far from perfect, so the development of new methods and the incorporation of these into the regulation will be essential.

Concretely, building this connection could look like:

- **More channels for filing complaints to the Commission and notifying authorities.**
  - An important way to monitor for and identify new sources of harm from AI systems is to allow more channels for filing complaints and raising concerns. One key channel would be allowing users to flag concerns directly. Currently the Act envisages users reporting to providers, who report to notified bodies, but other channels would enable the Commission to make better use of reported information. For example, for concerns meeting a particular bar of severity (as defined in the Act), there could be a channel for users to raise concerns directly with authorities. Other channels for raising concerns should include provisions for concerns and complaints to be raised by representatives of users and EU citizens, such as NGOs and unions; competing firms (as in EU competition law); and internal company whistle-blowers.
- **Ensuring that some Board members have expertise in AI monitoring, assessment, and horizon-scanning.**
  - If the Board is empowered to propose changes to the Regulation’s Annexes, this would allow insights from monitoring and progress in assessment to be incorporated into the classification and assessment of high-risk systems.
- **Maintaining an ‘AI safety incidents’ database at a European level.**
  - This would enable faster learning and feedback across the industry. It would also clarify scope and orientation and the potential need for regulatory revisions or

updates. This could easily be done by establishing an anonymised database, based on the serious incidents reported under Article 62.

- **EU institutions and Member States funding more work (within government and academia) on AI monitoring and assessment.**
    - Successful implementation of the regulation will depend on innovation in both methods to monitor the evolving AI risk landscape and methods for assessing the impacts of AI systems. EU institutions and Member States should explore ways to directly fund this work in order to ensure it is as aligned with the goals of the regulation as possible.
- **Ensuring that the Commission can update Annex VIII** (information to be submitted for the public database).
    - The public database will be an important way to attribute harms, seek redress, enable post-market monitoring, and identify new sources of harm from AI systems. It will be especially important for users, regulators, academia, and civil society. The database needs to be kept up to date in light of technical progress - in coming years new forms of information may become important. This could include, for example: parameter count; a measure of the computational resources used to develop, train, test and validate the AI system; or measures of hyperparameters, such as ‘temperature’. However as the Act is currently drafted, Annex VIII (information to be submitted for the public database) cannot currently be updated. This seems like an oversight - it is important that the Commission can keep this, like other Annexes, up to date. This could be easily fixed by adding to Article 60: “6. The Commission is empowered to adopt delegated acts in accordance with Article 73 for the purpose of updating Annex VIII in order to introduce additional required information that becomes necessary in light of technical progress.”, and adding “and Article 60(6)” to the lists in 2, 3 and 5 in Article 73.