Consultation on the European Commission's *White Paper on Artificial Intelligence: a European approach to excellence and trust*

Submission of the Centre for the Governance of AI,
Future of Humanity Institute, University of Oxford

June 2020

Table of Contents

Introduction

The Centre for the Governance of AI, part of the Future of Humanity Institute, University of Oxford, strives to help humanity capture the benefits while managing the risks of artificial intelligence (AI). We conduct research and advise decision-makers on some of the most important and neglected issues of AI governance, drawing on political science, international relations, computer science, economics, law, and philosophy.

In this response, we focus our analysis and recommendations on the proposed "ecosystem of trust" and associated international efforts. We believe these measures can mitigate the risks that this technology poses to the safety and rights of Europeans. Trustworthy technology also contributes to the long-term competitiveness of the European AI sector. Accidents and misuse would risk undermining the trust necessary for this industry to flourish.[1] There is evidence that a majority of Europeans already want AI technology to be carefully managed and believe a public policy response is required.[2] At the same time, any regulatory framework on AI needs to be proportionate and flexible. Without such a careful approach, we risk stifling the innovation needed to reap the benefits from further development and adoption of this technology.

[1] The public response to the Three Mile Island Nuclear Accident, for instance, contributed significantly to the subsequent slowdown of the U.S. nuclear power industry. See Hultman & Koomey (2013). "Three Mile Island: The driver of US nuclear power's decline?" Bulletin of the Atomic Scientists, 69 (3). DOI: 10.1177/0096340213485949.
[2] TNS opinion & social (2017). "Attitudes towards the impact of digitisation and automation on daily life." Special Eurobarometer 460. Retrieved from: https://data.europa.eu/euodp/en/data/dataset/S2160_87_1_460_ENG. Kantar (2019). "Europeans and Artificial Intelligence." Standard Eurobarometer 92. Retrieved from: https://ec.europa.eu/commfrontoffice/publicopinionmobile/index.cfm/Survey/getSurveyDetail/surveyKy/2255

Executive Summary

Regulatory scope. We support the proposed cumulative risk-based approach, which combines a risk assessment of different sectors with an assessment of intended uses since it allows for requirements that are proportionate to the risks posed by a given application. We make several recommendations to clarify the proposed system. We also suggest incorporating additional aspects into the risk assessment procedure to better account for risks from applications with large-scale effects.

- *Clarify the definition of risk as a function of the likelihood of a harmful scenario occurring and the severity of that harm.*
- *Clarify the assessment of multi-sector AI applications.*
- *Consider incorporating harm to public interests in the risk assessment.*
- *Consider incorporating the scale of use (number of users, frequency of use) of a given AI application into the risk assessment procedure.*

Type of requirements. We agree with the spirit of the proposed requirements. We make two recommendations to address particular failure modes of AI applications.

- *Consider adding specific robustness requirements for AI applications operating in tightly coupled systems prone to emergent behaviour patterns.*
- *Consider requiring disclosures of conflicts of interest between AI applications and their users.*

Compliance and enforcement. We are generally supportive of the proposed compliance and enforcement mechanisms. Our recommendations focus on ensuring the development of tools and expertise required for putting them into practice.

- *Support research on testing, evaluation, verification, and validation (VVT&E) of AI applications.*
- *Support research on the interpretability of AI applications.*
- *Support social science research on the impact and governance of AI systems.*

Governance. We agree with the need for a flexible regulatory framework and make recommendations for evaluation and monitoring efforts, intended to inform future adjustments.

- *Regularly review and amend the regulatory framework.*
- *Consider establishing a database for the sharing of information on AI incidents.*

International aspects. We support the cooperative stance of the Commission and make concrete recommendations for increased international collaborations.

- *Initiate and support international efforts for developing a shared understanding of the potential benefits and risks from AI.*
- *Facilitate exchange with global players on best practices for the assessments, testing, and regulation of AI applications.*
- *Contribute to the setting of global AI technology standards.*
- *Explore the possibility of establishing an international database for the sharing of information on AI incidents.*
- *Explore international partnerships for the proposed lighthouse research centre in Europe.*

Regulatory scope

We are supportive of a risk-based framework, ensuring "that the regulatory intervention is proportionate" (p. 17).[3] For the risk assessment, the two cumulative criteria proposed in the White Paper are reasonable. Allowing for exceptional instances that are to be considered "high-risk" as such is sensible.

We believe that a few aspects of the assessment procedure would benefit from clarifications, which we outline below. We also suggest incorporating additional factors into the risk assessment procedure to better account for risks from applications with large-scale effects.

Clarify the definition of risk as a function of the likelihood of a harmful scenario occurring and the severity of that harm. The Commission suggests that a sector is to be considered "high-risk", if, "given the characteristics of the activities typically undertaken, significant risks can be expected to occur" (p. 17). Furthermore, the specific uses have to be such that "significant risks are likely to arise" (p. 17). This approach does not seem to follow other existing risk assessment methodologies, which usually define risk in a given scenario as a function of the "combination of the probability of occurrence of a hazard generating harm in a given scenario and the severity of that harm."[4] Concretely, we are concerned that this approach does not sufficiently take into account potentially catastrophic risks involving AI systems since these are usually *unlikely* to occur. For instance, risks from the use of AI applications in nuclear power plants might not be "expected to occur" but should still be considered sufficiently serious to merit the designation as a "high-risk" application. Preventing and mitigating such risks through appropriate measures is particularly important since even just one occurrence would cause harm on a large scale.

[3] If not indicated otherwise, direct quotes are from: European Commission (2019). "White Paper: On Artificial Intelligence - A European approach to excellence and trust" (EN). Retrieved from: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
[4] European Commission (2015): "EU general risk assessment methodology" (EN), p. 4. Retrieved from: http://ec.europa.eu/DocsRoom/documents/17107/attachments/1/translations/

Clarify the assessment of multi-sector AI applications. The cumulative risk assessment procedure proposed by the Commission is appropriate for assessing narrow AI applications that can be situated within a single sector. However, we are concerned that this will result in legal uncertainties for systems that can be employed within multiple sectors (e.g., AI-enabled cybersecurity applications) or have general capabilities that extend across sectors (e.g., AI-enabled personal assistants).[5] The current phrasing of the risk assessment procedure is ambiguous with regard to how to handle such cases.[6] Such systems might evade regulatory scrutiny, even though they pose more risks than applications employed in just one sector.

Consider incorporating harm to public interests in the risk assessment. The Commission acknowledges that "the impact of AI systems should be considered not only from an individual perspective, but also from the perspective of society as a whole." (p. 2) The risk assessment in its current outline does not reflect this. It is our understanding that it is solely focused on legal entities. Some applications, however, will pose structural risks to public interests that would not necessarily be reflected in an assessment of the harm to persons or other legal subjects. For instance, AI-enabled recommendation algorithms on social media platforms could shape public discourse in ways that are harmful to our democratic system while not clearly harming the safety or violating the rights of any particular person. So the Commission could consider "protecting a wider range of 'subjects' [than consumers]. [...] These 'subjects' are normally the overall public interests covered by the relevant Union harmonisation legislation."[7] We suggest, however, to proceed carefully since attributing influence and harm in this domain is difficult.[8]

Consider incorporating the scale of use (number of users, frequency of use) of a given AI application into the risk assessment procedure. The White Paper suggests that the "assessment of the level of risk of a given use could be based on the impact on the affected parties" (p. 17). We generally support such an impact-based assessment. We are concerned, however, about an analysis that is exclusively based on the "risk of a given use." For determining whether oversight and regulation for a given application are appropriate (i.e., whether a system should be considered "high-risk"), what matters is not only the risk posed by a given use but also the overall risk posed by the application. This overall risk increases with the number of users and interactions of a given AI application. Applications interacting with millions of Europeans every day should face more scrutiny and oversight compared to applications interacting with hundreds of Europeans once a week, all else being equal. Such an approach

[5] While not necessarily common today, such systems may become more prevalent in the future. There are some indications that AI algorithms are becoming increasingly general in their capabilities (e.g., GPT-3, Alpha Zero, Agent57). It is likely that such advances will be reflected in future commercial AI systems as well. Personal assistant software can already perform tasks across a range of domains.

[6] There is no ambiguity only with regard to biometric identification applications and AI applications for recruitment processes. They are designated as exceptional instances of "high-risk" applications as such.

[7] European Commission (2015): "EU general risk assessment methodology" (EN), p. 9. Retrieved from: http://ec.europa.eu/DocsRoom/documents/17107/attachments/1/translations/

[8] For instance, many media reports had claimed that the YouTube recommendation algorithm suggested increasingly radical content to users of the platform. So far, however, this has not held up on closer inspection: Ledwich & Zaitsev (2019). "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization." arXiv:1912.11211.

could also reduce the regulatory burden faced by Small and Medium Enterprises (SME). We acknowledge that in some cases determining scale can be hard. Current risk assessment procedures, however, require similarly difficult estimates of the probability that a given harmful scenario will occur. We believe that reasonable estimates can be made. These can be adjusted as we learn more about the development of specific technologies.

Types of requirements

We agree with the spirit of the requirements laid out in Section D of the White Paper. We make two recommendations to address particular failure modes of AI applications.

Consider adding specific robustness requirements for AI applications operating in tightly coupled systems prone to emergent behaviour patterns. Such emergent patterns can cause harmful outcomes, in a manner not easily foreseen by consideration of any one AI application in isolation.[9] In the 2010 “flash crash” of the U.S. stock market, for instance, such interactions among trading algorithms led to a sudden trillion-dollar decline in U.S. financial market value.[10] The markets quickly returned to pre-crash levels, partly as a result of mandatory “circuit breakers”, a requirement that the Securities and Exchange Commission had introduced after the stock market crash of Black Monday in 1987. Such failures, however, might be extremely harmful and irreversible in other sectors, such as defence and cybersecurity.[11] Requiring applications to be sufficiently robust in this respect could prevent such outcomes to some extent. For instance, testing of such systems should not take place in isolation but mirror the tight coupling of the environment they will be employed in.

Consider requiring disclosures of conflicts of interest between AI applications and their users. Users increasingly consult AI applications for information, advice, and recommendations (e.g., via personal assistants, search results, or user feeds). They may reasonably expect that algorithms providing advice or recommendations are aligned with their best interests. Developers, however, are often incentivised to align the algorithm with their commercial interests, which track the best interests of their clients only imperfectly.[12] For instance, companies may have their virtual assistants promote their own products, even if a better product was available for the user. Given the strong information asymmetry inherent in the use of such applications, users on their own will be in a poor position to identify such misalignment. Requiring the disclosures of such conflicts of interest will empower users to make more

[9] Maas (2018). “Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment.” AAAI/ACM Conference. DOI: 10.1145/3278721.3278766.
[10] U.S. Commodity Futures Trading Commission, and U.S. Securities & Exchange Commission (2010). “Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues.” Retrieved from: https://www.sec.gov/news/studies/2010/marketevents-report.pdf
[11] Scharre (2018). “Army of None: Autonomous Weapons and the Future of War.” W. W. Norton & Company. Feldman, Dant & Massey (2019). “Integrating Artificial Intelligence into Weapon Systems.” arXiv:1905.03899v1.
[12] Aguirre, Dempsey, Surden & Reiner (2020). “AI loyalty: A New Paradigm for Aligning Stakeholder Interests.” U of Colorado Law Legal Studies Research Paper No. 20-18. Retrieved from: http://dx.doi.org/10.2139/ssrn.3560653

informed decisions about their engagement with such applications. Such requirements may involve disclosing the basic business model and types of revenue generated by a given application. They may also include labelling advertisements and sponsored content or results to the users of a given application.

Compliance and enforcement

We support an independent conformity assessment of "high-risk" applications prior to market entry to ensure that they meet the requirements set out in the regulatory framework. Self-assessment by businesses themselves would not be sufficient or fair, given the potentially catastrophic risks included in the "high-risk" classification and the burden self-assessments would impose on SME.[13] Additional controls after applications have already entered the market are particularly important for "certain AI systems evolve and learn from experience" (p. 23).

Our recommendations focus on ensuring the development of tools and expertise required for effective compliance and enforcement.

Support research on testing, evaluation, verification, and validation (VVT&E) of AI applications. Conformity assessments will require testing, evaluation, verification, and validation (VVT&E) of AI applications, especially for verifying conformity with robustness requirements. Unfortunately, the "current state of AI VVT&E is nowhere close to ensuring the performance and safety of AI applications, particularly where safety-critical systems are concerned."[14] Methods used for ensuring the safety and reliability of traditional software, e.g., formal verification, can currently not be applied to machine learning systems. Advancing this field will be crucial for establishing an effective assessment scheme. We suggest the Commission support such work through its Framework Programmes. Test centres could also be given support to conduct research where necessary.

Support research on the interpretability of AI applications.[15] Progress on the interpretability of machine learning systems would likely contribute to a well-functioning liability regime since it would make it easier for claimants to establish a cause-and-effect relationship between model and harm. It would also improve our predictions of the output of AI systems ex-ante, which would be helpful for developers designing more trustworthy applications as well as making risk assessments and conformity assessments more reliable. Unfortunately, "there is very little consensus on what interpretable machine learning is and how it should be measured."[16] Establishing clear definitions and benchmarks for interpretable AI applications will be important

[13] If the risk classification scheme was more fine-grained, self-assessment might be appropriate for some tiers.
[14] Tarraf et al. (2019): "The Department of Defense Posture for Artificial Intelligence." RAND report, p. 36. Retrieved from: https://doi.org/10.7249/RR4229
[15] For an interview of interpretability research: Gilpin et al. (2019). "Explaining Explanations: An Overview of Interpretability of Machine Learning." 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018). arXiv:1806.00069.
[16] Doshi-Velez, Kim (2017). "Towards A Rigorous Science of Interpretable Machine Learning." arXiv:1702.08608.

for enabling progress in the field. We suggest the Commission support such work through its Framework Programmes.[17]

Support social science research on the impact and governance of AI systems. We agree with the Commission that "authorities should be in a position to investigate individual cases, but also to assess the impact on society" (p. 23). Such research into the social effects of AI systems and how to address them through appropriate governance systems is still in its infancy. Advancing research in this field would be important for informing the work of the relevant national and European authorities as well as guide the further development of the regulatory framework. We suggest the Commission support such work through its Framework Programmes.

Governance

AI is a rapidly evolving field. In light of this, we agree with the Commission that "the regulatory framework must leave room to cater for further developments" (p. 10), instead of trying to create a regulatory framework that anticipates all such eventualities.[18] Addressing future developments will require thorough evaluation and monitoring efforts, which is the focus of our recommendations.

Regularly review and amend the regulatory framework. We trust the Commission to set up frequent and timely evaluations as well as appropriate monitoring mechanisms, following its guidance as laid out in the *Better Regulation Guidelines*.[19] We are supportive of the stakeholder consultation procedures outlined in Section H of the White Paper.

Consider establishing a database for the sharing of information on AI incidents. A central repository of "AI incidents", i.e., instances of undesired or unexpected and (potentially) harmful behaviour by an AI application, would improve the implementation and further development of the regulatory framework. The appropriate national authorities and independent testing centres could build up shared institutional knowledge of common failure modes. The Commission would also be in a better position to adjust the scope and requirements of the framework. Such a database has been proposed by a broad coalition of researchers and the Partnership on AI has already launched such a database.[20] Widespread use would be crucial for its success, but businesses might be wary of submitting incident reports for fear of reputational or other costs.

---

[17] Brundage, Avin, Wang, Belfield, Krueger, et al. (2020). "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." arXiv:2004.07213.
[18] Bostrom, Dafoe & Flynn (2016). "Policy Desiderata for Superintelligent AI: A Vector Field Approach." Forthcoming in Liao, S.M. (ed.): Ethics of Artificial Intelligence (Oxford University Press). Retrieved from: https://pdfs.semanticscholar.org/9601/74bf6c840bc036ca7c621e9cda20634a51ff.pdf
[19] European Commission (2017). "Better Regulation Guidelines." SWD (2017) 350. Retrieved from: https://ec.europa.eu/info/sites/info/files/better-regulation-guidelines.pdf
[20] Brundage, Avin, Wang, Belfield, Krueger, et al. (2020). "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." arXiv:2004.07213. For the database, see Partnership on AI. "AI Incidents Database." Available at: http://aiid.partnershiponai.org/

Several measures could be taken to address such concerns, e.g., by ensuring anonymity, security, and privacy. Submitting incident reports could be further incentivised, e.g., through regulatory requirements or monetary incentives for developers, users, or third parties ("bounties").[21] Similar models for such information-sharing arrangements have been set up in other areas. As part of European pharmacovigilance efforts, the European Medicines Agency, for instance, established a central electronic repository for periodic safety update reports by firms in the pharmaceutical industry.[22] The Commission could apply best practices from similar such registries, where applicable.

International aspects

All nations will face both competitive and cooperative pressures in the international environment with regard to AI technology. Competitive dynamics arise out of national interests for relative economic and military advantage, favouring speed and innovation over safety and stability. Cooperative dynamics arise out of a shared interest in innovation, growth, safe technology, and international stability. Without trust and coordination, however, cooperative dynamics will be difficult to sustain. We, therefore, encourage the Commission to continue its commitment "to cooperate with like-minded countries, but also with global players, on AI, based on an approach based on EU rules and values" (p. 8). In our recommendations, we focus on concrete areas of collaboration.

Initiate and support international efforts for developing a shared understanding of potential risks from AI. There appear to be differences in the conception and terminology of AI risks and safety among different countries.[23] Arriving at shared definitions and understanding could prevent miscommunication and lay groundwork for further collaboration in the area of trustworthy AI. Convening relevant (technical) experts and policymakers from global players could be the first step. The Pugwash Conferences could serve as a model. These regular, informal conventions with participants from the United States and the Soviet Union during the Cold War very likely contributed significantly to several international arms control agreements.[24]

Facilitate exchange with global players on best practices for the assessments, testing, and regulation of AI applications. Policymakers and auditors across the world face similar regulatory and technical challenges for ensuring the trustworthiness of AI applications. Mutual

[21] Brundage, Avin, Wang, Belfield, Krueger, et al. (2020). "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." arXiv:2004.07213
[22] European Commission (2016). "Pharmacovigilance related activities of Member States and the European Medicines Agency concerning medicinal products for human use (2012 – 2014)." COM(2016) 498 final. Retrieved from: https://ec.europa.eu/health//sites/health/files/files/pharmacovigilance/pharmacovigilance-report-2012-2014_en.pdf
[23] Imbrie, Kania (2019). "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement." Center for Security and Emerging Technology Policy Brief. Retrieved from: https://cset.georgetown.edu/research/ai-safety-security-and-stability-among-great-powers-options-challenges-and-lessons-learned-for-pragmatic-engagement/
[24] Rubinson (2019). "Pugwash Literature Review." Urban Institute. Retrieved from: https://www.urban.org/sites/default/files/pugwash_literature_review.pdf

learning could improve the development and implementation of regulatory frameworks as well as contribute to regulatory convergence. The Commission could facilitate the convening of technical experts from regulatory authorities, standards organisations as well as independent auditors and testing centres. They could work towards the development of global standards for the assessment and certification of specific AI applications.

Contribute to the setting of global AI technology standards. International standards organisations like the ISO, the IEC, and the ITU are important multilateral fora in the context of AI because they include all global players.[25] The White Paper does not sufficiently emphasise the constructive and influential role that the Europe Union could play in these international organisations through the respective cooperation agreements of CEN and CENELEC.[26] With its regulatory and technical expertise, European influence could contribute to more robust standards, which are also in line with European values.

Explore the possibility of establishing an international database for the sharing of information on AI incidents (see "Governance" section). Bringing on additional international partners for this project could improve the safe development and deployment of AI systems around the world. The International Atomic Energy Agency, for instance, maintains a similar database for incident reports related to the operation of nuclear power plants.[27]

Explore international partnerships for the proposed lighthouse research centre in Europe. As a result of projects like ITER and CERN, the EU has experience bringing together different, even competing, nations for large-scale research projects for the benefit of humanity. Focusing the lighthouse research centre on shared interests like trustworthy AI (e.g., research on interpretability or privacy-preserving machine learning) would distribute the costs for providing a public good and ensure that the collaboration is in the interest of a broad international coalition. The resulting increase in funding could also establish the project as a leading research lab in the world.

[25] Cihon (2019). "Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development." Technical Report, Centre for the Governance of AI. Retrieved from: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf
[26] Büthe, Mattli (2011). "The New Global Rulers: The Privatization of Regulation in the World Economy" (p. 138-9). Princeton University Press.
[27] International Atomic Energy Agency. "Incident Reporting Systems for Nuclear Installations." Available at: https://www.iaea.org/resources/databases/irsni