DeepMind

DeepMind response to the Artificial Intelligence Act

Overview

The development of artificial intelligence (AI) raises important and complex questions; its impact on society and on all our lives is not something that should be left to chance. We believe AI's extraordinary potential will only be realised if its development and deployment upholds appropriate ethical standards and is purposefully directed towards benefitting society.

DeepMind welcomes the European Commission's vision to promote responsible, human-centric and trustworthy AI, and the proportionate and risk-based approach it has taken in the Artificial Intelligence Act (AIA). The AIA has an opportunity to both foster and regulate technology development, which are crucial to ensuring innovation, safety, and competitiveness. AI systems and associated risks are likely to evolve, and a growing body of research may address these risks along the way. This is why it will be essential to set up appropriate mechanisms to inform and update the regulation, where needed over time, including through transparent reviews of the categories of high-risk AI. To promote the uptake of AI and create an ecosystem of excellence in the EU, clarity on the parameters of the AIA – particularly as they pertain to early stage research – would be welcome, and we would encourage a greater referencing of the Coordinated Plan on AI in the AIA.

We believe that the responsibilities allocated to different actors in the current text do not fully reflect the complexities of the AI ecosystem. The AIA proposes requirements for high risk AI systems and allocates those to a set of actors, with the main burden resting on "providers" of AI systems. However, the way AI systems are initially developed, then revised, shared, and integrated by different actors in practice leads to many different scenarios that are difficult to map to the AIA. This allocation of responsibilities needs to be particularly nuanced when considering general purpose AI systems.

Mandatory requirements for AI systems also need to be technically feasible. While we support the spirit of the requirements, some of the current phrasing raises doubt as to their technical feasibility and would benefit from refinements and scope clarifications. We offer

particular suggestions for requirements regarding datasets, record keeping, access to source code, and human oversight.

A robust regulatory AI framework should help drive AI community energy, investment and research toward important technical challenges. The provisions around accuracy, robustness and cybersecurity are essential and we hope some of the governance processes established with the AIA, such as sandboxes or the European AI Board, would also prioritise AI safety in their mandate and operational priorities.

Finally, we support greater cooperation between governments to establish a harmonized global approach to AI governance and avoid regulatory fragmentation.

We look forward to contributing further to these important discussions and welcome the opportunity to answer any questions on our feedback.

About DeepMind[1]

DeepMind is a scientific discovery company, committed to 'solving intelligence' to advance science and humanity. This requires a diverse and interdisciplinary team working closely together – from scientists and designers, to engineers and ethicists – to pioneer the development of advanced artificial intelligence. AI has the potential to enrich the lives of billions and improve our understanding of the universe. Ultimately we hope that new scientific breakthroughs, driven by innovations in machine learning, can make the crucial difference in helping us prosper in an increasingly complex world, and respond to tough challenges such as climate change and tackling diseases.

With our deep learning system AlphaFold, for instance, we brought together experts from the fields of structural biology, physics, and machine learning to apply cutting-edge techniques to predict the 3D structure of a protein, based solely on its genetic sequence and showed how artificial intelligence research can drive and accelerate new scientific discoveries. We are excited for resources such as the AlphaFold Protein Structure Database, which we recently released in partnership with the European Molecular Biology Laboratory's (EMBL) European Bioinformatics Institute (EBI), to herald a new age of AI-powered scientific breakthroughs.

AI can bring extraordinary benefits to society, but only if it is built and used responsibly. For DeepMind, responsibility means aligning our research with society. We view responsible AI

[1] We share these comments on behalf of DeepMind, and not on behalf of Google or any other entity in Alphabet, Inc.

as an ongoing process of ensuring our research and engineering are informed by the values, needs and expectations of society. In practice, this means: (1) making sure our research addresses major scientific and social challenges; (2) anticipating and mitigating potential harms; and (3) engaging with the wider world, its complexities, challenges and possibilities.

We welcome the European Commission's vision to promote responsible, human-centric and trustworthy AI through effective investment and regulation, and the opportunity to provide feedback on the Artificial Intelligence Act.

The importance of a proportionate, risk-based approach

We are supportive of the regulation's proportionate approach and focus on high-risk AI systems. Routine, low-risk AI applications should be held to different standards to those in high-risk settings, and certain uses of AI should be prohibited - especially when they breach fundamental rights and values. By focusing on understanding the specific contexts and ways in which AI is likely to be deployed, and the potential impact it could have on citizens, rules on AI can help foster trust, direct investment toward greater safety and reliability in AI systems, and help society better prepare for increasingly advanced AI. At the same time, it is important for the rules to be clear and focused, in order not to stifle the ability of new entrants to innovate nor deprive society of the many potential benefits of AI.

Driving excellence and innovation in the AI ecosystem must happen in parallel to regulating risky uses of AI. This is why we welcomed the publication of the Coordinated Plan on AI, which contains a series of crucial recommendations to improve the EU's AI ecosystem of excellence. We believe a stronger link between the two documents would better reflect a bold EU vision to promote AI research and safe AI development, and we recommend the AIA makes greater reference to the Coordinated Plan on AI. A cornerstone will be ensuring that new rules do not discourage much needed investment in AI research. The huge potential for AI to advance scientific discovery will require a significant expansion in fundamental research, much of it curiosity-driven and speculative. It is vital that the EU's approach to regulation continues to encourage ambitious blue-skies work.

Flexibility and adaptability to keep pace with innovation

AI is a complex and evolving technology; we are therefore fully supportive of the European Commission's intention to make this regulation flexible and adaptable. A key element of this will be to ensure there are adequate mechanisms to inform and update the regulation, where needed, over time.

This is particularly relevant when it comes to the risk classification. The pace of AI development is incredibly rapid, and regulators should be prepared for that acceleration to increase over time. We believe the advancing state of AI will help us make progress on challenges that are currently very difficult – including the challenges of bias, explainability, and safety that underpin the risk classification in this proposal. On the other hand, increasingly advanced, general purpose, and distributed AI may lead to risks that are only likely to manifest in the longer term. The list of high-risk AI should allow for modifications to the areas listed in Annex III and not just specific uses within the existing categories. It needs to allow for the possibility that some of those uses might become *less* risky over time, and therefore worthy of being dropped from the list. Likewise, it should be possible to add entire new categories of risk, including ones that may not be currently contemplated, in light of the evolution of technology and certain dynamics that may emerge from specific characteristics of complex AI systems (such as having many AI systems interacting with one another) and the fields in which they may be deployed.

This process to update the list of high risk AI systems will need to be robust, systematic, and transparent. Unpredictable, haphazard changes to the list could sow confusion and impede on AI development. The process should rather incorporate opportunities for multi-stakeholder feedback, including from industry and civil society, at regular intervals that offer the ability to monitor and plan the development of AI systems accordingly.

Clarifying scope and responsibilities

As the AIA puts forth an extensive new mapping of actors and responsibilities across the AI ecosystem, its success in fulfilling its vision for proportionality hinges on having definitions that are both clear and able to provide a balance of certainty and flexibility over time. This is particularly important given the very expansive breadth of the general definition of "AI systems" in Annex I. As such, we offer suggestions for particular areas that would benefit from greater nuance and clarification, especially to ensure that research continues to flourish; and to set the right balance of responsibilities between providers, distributors, users, third-parties and importers.

Promoting important AI research

The AIA would benefit from greater clarification that it would not unnecessarily constrain AI research, including the dissemination of knowledge via publications, and instead remains focussed on the real-world impact of AI applications. As and when advances in research suggest a need to revise the high risk classification for applications, there should be a dynamic process to facilitate this. While Recital 16 notes that research – in the context of

the "*use of certain AI systems intended to distort human behaviour*" – should not be stifled by the prohibition on unacceptable uses, we would encourage including in an operative clause a more general statement that research as it relates to all *"AI systems,"* and assuming *"such research does not amount to use of [an] AI system in human-machine relations that exposes natural persons to harm,"* would continue to be encouraged.

Balance of responsibilities

The AI ecosystem is complex, and as research may transition into a specific, operational AI system, it is important to consider the way different actors interact, and how responsibilities ought to be allocated in different scenarios. AI applications may contain many different AI systems within them. These systems may themselves be developed by single entities or as a result of collaborations and may involve various forms of third-party integration. The obligations for high-risk systems outlined in the AIA are largely deployment-dependent, involving questions of interactions with users and ongoing monitoring and documentation. We believe every actor in the AI ecosystem should contribute to responsible development and dissemination of AI in ways that are proportionate to their position in the ecosystem. As a result, we suggest that the obligations under the AIA be clearly tailored to both (1) an actor's practical ability to comply and (2) circumstances most likely to lead to the harm that the AIA aims to prevent.

As software gets integrated into operational AI systems, it is important to reflect on how responsibilities are best allocated. Open-source software is an important part of the research ecosystem, helping advance knowledge and drive scientific collaboration, and often released alongside research publications as a scientific norm.[2] Carefully tailored measures are needed to avoid impeding on this general practice. A significant part of innovation in AI is also based on the increasing integration into applications of off-the-shelf AI, including in the form of APIs and machine learning toolkits. As more technology is made available in this way, we will see increasing situations where entities may deploy AI systems in ways that do not modify the underlying system, but are nonetheless beyond the awareness or control of the original AI developer. Particularly where that technology may be more general-purpose in nature, there may be circumstances where AI technology gets integrated into a high risk system, yet the original developer of the system will not be in a position to fulfill the compliance obligations outlined in the AIA.

[2] For example, in July, we released a peer-reviewed paper in *Nature* explaining how AlphaFold works, and made available open source code. This is a key way that we hope for the scientific community to understand and build on our system, and drive further breakthroughs in the field.

Specifically, we recommend that the role of the 'provider' be more clearly defined in Article 3(2) to be *"an entity that is making a high risk AI system available for use by end users,"* and a clear inclusion in Title III that only 'providers,' or other entities that assume the key characteristics of the 'provider' role, are subject to the mandatory obligations for high risk AI systems. We also recommend clarifying (for example, in Article 28) that an entity may assume the role of such a 'provider' even where there has not been substantial modification of an underlying AI system.

In these cases, it may be more appropriate, for example, for the original developer to share adequate upfront information alongside their AI system, which other parties could use as needed as a basis to fulfill the mandatory requirements, as they apply to the operational AI system.

Setting workable compliance obligations

We welcome the fact that mandatory requirements are focused on AI systems considered to carry the greatest risks. However, while we support the spirit of the requirements, some of the current phrasing would not be technically feasible. We suggest the following clarifications:

- *Data governance:* The current draft of the regulation (Article 10.3) states that "training and testing data sets shall be relevant, representative, free of errors and complete"; which sets a standard that isn't feasible. In practice, it is unlikely to be the case that a data set will ever be completely "free of errors." Furthermore, even narrow datasets can yield results that are appropriate for a particular use if the AI system is designed with the right constraints. It is more meaningful for "formulation of relevant assumptions" to be incorporated into data management practices as appropriate, as called for in the text, and to request that "adequate efforts should be made to ensure that training and testing datasets are sufficiently relevant, representative and robust in view of the intended purpose of the system," recalling the guidance in Recital 44.

- *Record keeping*. While we agree that in principle keeping records and retaining data sets is beneficial, and a practice companies often already do for their internal record-keeping, we caution against the requirements as drafted in the AIA. Audit logs may have some merits, for example with respect to explainability, but they can also create privacy risks, especially if they contain personal information. Instead traceability could be obtained in more efficient ways, and retaining logs in a proportionate manner for a limited time could be a workable alternative.

- *Access to source code*. The requirement that "market surveillance authorities shall be granted access to the source code of the AI system" if they have some grounds to suspect non-compliance should be replaced by more efficient methods for verifying the performance of an AI system, which would not put at risk providers' intellectual property rights. We believe this should be replaced with an ability to request that AI providers and deployers "equip market surveillance authorities with the necessary information – including on system inputs, outputs, and/or training methodology – to carry out robust testing where it is necessary to confirm compliance."

- *Human oversight*. The AIA requirement for a human to "fully understand the capacities and limitations" of a high risk system (Article 14.4(a)) would be unworkable for certain AI systems at the moment. Indeed, the transformative potential for AI to advance humanity is largely grounded in the way it can help us make sense of information and amplify human capabilities beyond what would otherwise be possible – including in ways that might be novel. This applies even in high-risk situations, where use of AI can help drive innovation and safer outcomes that would not be possible without its use. It is, however, crucial for the AI system in question to allow meaningful and effective human oversight in these scenarios. We believe replacing "fully understand" by "adequately understand" would meet this policy objective in a more balanced way. Finally, while "full understanding" may not be possible at present, technological developments will need to be considered over time. Future reviews of the AIA may need to account for changes in our ability to understand and explain AI systems.

- *Codes of conduct*. We were glad to see the AIA encourage companies which are developing AI systems with minimal risk to develop voluntary commitments related to, for instance, environmental sustainability or diversity of development teams, which could prompt companies to strengthen their internal responsible AI governance processes. However, the current phrasing leaves it unclear what is expected of companies when it comes to the codes of conduct, as it is possible that all AI providers feel pressured to mirror the requirements for high-risk AI systems noted in the AIA, creating an unnecessary administrative burden and cost, and defeating the proportionate objective of the regulation. Instead, companies should be encouraged to develop internal standards and a space should be created for companies to exchange best practices through organisations such as the Partnership for AI, and to coordinate through standard-setting organisations the development of norms to improve the safety and ethics of their AI development.

Prioritising innovation in AI safety

We are supportive of the AIA requirement for high-risk AI systems to be "designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle." We appreciate the decision to leave the precise technical solutions to standards or other technical specifications. We encourage the Commission to consider how to foster industry, civil society and governmental discussions on improving AI safety, potentially through supporting the creation of separate fora for these actors to share their approach in a way that could feed into future guidance and into AI safety discussions that also need to take place within the European AI Board.

Over the long term, and as technology advances, a robust regulatory AI framework should help drive AI community energy, investment and research toward solving important technical challenges. Despite their importance, technical and sociotechnical research in areas such as robustness, reliability, safety, bias, explainability and privacy remain significantly under-resourced. This is a crucial gap, as these desired attributes of AI systems are also open research problems. For example, there are many promising research approaches that could help to better 'explain' the predictions or behaviour of complex AI models, for example via language models and conversational agents, as well as approaches to evaluate how 'useful' these explanations are to humans in different contexts - something that rarely occurs at present.

Increased investment in R&D that advances responsible AI should be a priority and we hope these provisions within the AIA will increase companies and researchers' efforts towards responsible AI. Policymakers also have a critical role in focusing attention on safe and responsible AI; their leadership in developing and upholding rules and standards will be key to social acceptance and public confidence in AI.

This is why we were pleased to see the provisions around establishing regulatory sandboxes in the AIA. Sandboxes could help provide a space for companies to innovate on long-term AI systems and foster high potential technologies, in a way that supports EU competitiveness and helps policymakers stay abreast of the latest technological developments. A focus within the sandboxes on AI safety could be another way to support the development of safe and responsible AI in the EU. The experience from the Financial Conduct Authority in the UK showed sandboxes could be an efficient way for start-ups to test their products in a regulated space and innovate, while also allowing for international cooperation between different sandboxes. In that regard, we hope that different sandboxes within the EU will be

able to share best practices and increase capacity within Member States, but also look at opportunities for collaboration with third countries. Regulatory sandboxes will also need to have sufficient capacity to cater to what could be significant demand.

Promoting a harmonized global approach to AI governance

We welcome the regulation's aim to reduce fragmentation within the EU when it comes to AI. Europe is home to a vibrant community of academic, industry and public organisations driving forward scientific research on AI. However, many of these efforts are currently fragmented and siloed, with limited coordination at the EU level. The AIA will undoubtedly increase this coordination between EU countries. It will be important to ensure that there aren't large discrepancies in the way Member States implement the AIA and we encourage national market authorities to work together to share best practices and coordinate their approach.

We welcome the establishment of a European AI Board, and its potential to foster more substantive discussions, to increase governments' expertise and capacity in AI, and to further reduce fragmentation between EU countries. We believe such an institution will help drive forward the EU's leadership in AI and hope that there will be opportunities for non-governmental actors to participate or be regularly informed of the work and decisions made by the European AI Board. We also welcomed the announcement by the European Commission that it is considering creating an expert group under the Board, modelled after the High Level Expert Group on AI, which could be an avenue for the private sector and civil society to engage with governments. The expert group could also be consulted when changes are made to the regulation, such as the list of high risk AI systems in Annex III. The benefits of regular and formal input from non-governmental actors can help further build a shared understanding of the potential and risks of AI, and exchanging of best practices. We would also encourage linking the European AI Board and the expert group with the existing actors in the EU AI ecosystem, such as the European AI Alliance.

Finally, we also need to acknowledge the impact the AIA will have globally as a first of its kind, and encourage greater cooperation between governments to establish a harmonized global approach to AI governance. We believe AI should not be a zero sum game for global powers. Its opportunities and risks will not be limited to any one country and governments should work together to counter unhelpful competitive race dynamics, promoting collaboration for its responsible development and use globally, which will be essential to build citizen trust globally. Greater international cooperation on AI can also help guard against malign uses of the technology by establishing and upholding standards, and ensuring it is used in line with democratic values, human rights, and the rule of law.

The EU has driven early progress on governance and cooperation internationally, and this can be built upon through its founding membership of the Global Partnership on AI, given the breadth of the countries involved and their common agreement on the OECD's AI principles. We're very encouraged to see the creation of more fora for international collaboration on AI, such as the EU-US Trade and Technology Council and hope it will lead to more interoperability between the EU and the US on AI safety and standards, and hope more convergence will be sought with other countries developing AI assurance systems such as the UK.