The Amsterdam Paper

Recommendations for the technical finalisation of the regulation of GPAI in the AI Act by the participants of the joint Media & Democracy Lab and Algorithm Watch legal design workshop

Helberger, N.; Naudts, L.P.A.; Piasecki, S.; Aszodi, Nikolett; Berendt, Bettina ; Brown, Ian; van Daalen, O.L.; Diakopoulos, N.A.; de Jonge, Tim ; Elmer, Christina; Helming, Clara; Iwanska, Karolina ; Keller, P.; Kreuter, Frauke; Obrecht, Liliane ; Mueller, Angela; Pannatier, Estelle; Oberski, Daniel; Quintais, J.P.; Spielkamp, Matthias; Tarkowski, Alek; Vieth-Ditlmann, Kilian; Weerts, Sophie; Zuiderveen Borgesius, F.J.

Publication date
2024

Document Version
Final published version

License
Unspecified

The Amsterdam Paper: Recommendations for the technical finalisation of the regulation of GPAI in the AI Act by the participants of the joint Media & Democracy Lab and Algorithm Watch legal design workshop[1]

Editors and contributors: Dr. Natali Helberger (Distinguished University Professor for Law and Digital Technology, University of Amsterdam, AI, Media & Democracy Lab), Dr. Laurens Naudts (Postdoctoral Researcher in Law, AI, Media & Democracy Lab), Dr. Stanislaw Piasecki (Postdoctoral Researcher in Law, University of Amsterdam, AI, Media & Democracy Lab)

Contributors:
Nikolett Aszódi (Policy & Advocacy Manager, AlgorithmWatch), Dr. Bettina Berendt (Professor for Internet and Society, TU Berlin), Dr. Ian Brown (Consultant; Visiting Professor at the Centre for Technology and Society at Fundação Getulio Vargas Law School, Rio de Janeiro), Dr. Ot van Daalen (Lawyer; Lecturer and Researcher in Information Law, University of Amsterdam), Dr. Nick Diakopoulos (Professor in Communication Studies and Computer Science (by courtesy) at Northwestern University), Tim de Jonge (PhD candidate, Radboud University), Christina Elmer (Professor for Digital Journalism/Datajournalism, University of Dortmund), Clara Helming (Senior Policy & Advocacy Manager, AlgorithmWatch), Karolina Iwańska (Digital Civic Space Advisor, European Center for Not-for-Profit Law), Paul Keller (Director of Policy, Open Future), Dr. Frauke Kreuter (Professor for Statistics and Data Science, LMU Munich), Liliane Obrecht (PhD Candidate in Law, University of Basel), Dr. des. Angela Müller (Head of Policy & Advocacy, AlgorithmWatch), Dr. Daniel Oberski (Professor in Health Data Science, Utrecht University), Estelle Pannatier (Policy & Advocacy Manager, AlgorithmWatch CH), Dr. João Quintais (Assistant Professor in Information Law, University of Amsterdam), Matthias Spielkamp (Founder & Executive Director, AlgorithmWatch), Alex Tarkowski
(Director of Strategy, Open Future), Kilian Vieth-Ditlmann (Deputy Team Lead Policy & Advocacy, AlgorithmWatch), Dr. Sophie Weerts (Associate
Professor in Public Law, University of Lausanne), Dr. Frederik Zuiderveen Borgesius (Professor of ICT and Law, Radboud University)

1. Introduction

The AI Act will become a reality. The Act's main objectives are to determine harmonised standards that high-risk AI systems would need to comply with (for example, when AI is used in hiring processes, the justice system or in the educational context).[2] However, the recent popularisation of general purpose AI (GPAI) across the world has led the EU legislator to add

[1] This long read does not necessarily reflect the opinions of all participants. The short version agreed upon by all workshop participants can be consulted here: https://algorithmwatch.org/en/wp-content/uploads/2023/09/PolicyBrief_GPAI_AW-updated1509-02.pdf.

[2] Michael Veale, Kira Matus and Robert Gorwa, 'AI and Global Governance: Modalities, Rationales, Tensions' (2023) 19 Annual Review of Law and Social Science 13 <https://www.annualreviews.org/loi/lawsocsci> accessed 7 November 2023.

a new dimension to the regulation in order to reflect these developments. The AIA will include definitions, transparency requirements and mandatory risk evaluations for at least certain general-purpose AI systems (GPAI – which includes generative AI systems such as ChatGPT or Bard) as well as various other provisions related to those recently released technologies.[3]

Once the AI Act has been agreed on, the work on transforming the – often abstract and general – concepts and legal requirements into a effective governance framework will begin. This will be done through implementing acts by the European Commission, technical standards from the European standardisation bodies, guidance from the AI Office as well as self-regulation, codes of conduct and the Terms of Use of technology companies. In practice, while the AI Act will lay down the (high-level) legal requirements, the latter additional guidance documents, codes, private rules and standardisation efforts will have an important role in the actual operationalisation of the law.

The vivid discussions around possible approaches to govern GPAI over the past months in and outside Europe have demonstrated the extent of complexity of this task, the level of uncertainty and lack of ready-made and tested solutions, but also the value of joining forces and bringing together the expertise of regulators, civil society and researchers from various disciplines. In this spirit, the goal of this document is to summarise the main insights from a workshop that brought together a group of academics and civil society representatives in Amsterdam. The objective of the workshop was to share and learn from each other's insights and expertise and jointly brainstorm on the problem of how to govern GPAI and what are possible routes for the AIA. The workshop was held in July 2023 and co-organised by AlgorithmWatch and the AI, Media & Democracy Lab of the University of Amsterdam. We publish the report in the hope that it can be a useful source of information and critical reflection for finalising and operationalising the European governance approach to GPAI. Given their specific focus on GPAI, the recommendations in this report are not comprehensive but should be regarded as complementary to other recommendations on the AIA. Please find below a summary of the main topics discussed in the report.

Firstly, this report discusses the **mandatory risk evaluations** introduced by the AI Act in the context of GPAI systems. The AI Act's role as an instrument of democratic oversight is analysed and complemented by some suggestions on how to further improve it (2).

Secondly, this reports reflects on **closing the accountability gap,** highlighting among others the issue of contractual fairness, new cooperation obligations, new obligations for deployers, as well as more specific provisions related to open source models (3).

Thirdly, GAI systems will need to comply with certain **transparency requirements** (Art. 52 AI Act). Art. 52 is one of the few provisions providing concrete entitlements to individuals affected by AI systems. This report begins by introducing the concept of transparency (summarising the current approach taken in the AI Act) and underlining the importance of its reflection, observability and actionability components. Subsequently, it identifies limitations of information obligations contained in Art. 52 AI Act, focusing both on the substance and modalities of information provision. The report provides recommendations on how to

[3] Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review <https://policyreview.info/essay/chatgpt-and-ai-act> accessed 7 November 2023.

enhance transparency for affected persons, improve contestation mechanisms (for individuals and collectively) and support external accountability purposes (including risk assessments) (4).

Fourthly, another essential evaluation component of any artificial intelligence system should be its **environmental impact**. Our recommendations underline that the AI Act must not miss the chance to integrate this crucial topic in order to ensure the sustainability of GAI and prevent excessive negative impacts on the environment (5).

Fifthly, various revelations regarding the exploitation of **data labelling GAI workers** underscore how crucial it is to provide them with rights and effective enforcement capabilities in the AI Act in order to ensure that they are adequately protected and their fundamental rights preserved (6).[4]

Sixthly, the proposed AI Act will also contain new **copyright-related** provisions requiring transparency about training data. How feasible are these provisions? Are there any other copyright-related considerations that should be included within the AI ACT? This report explores those questions and provides relevant answers (7).

Seventhly, while certain GAI-related **data protection** issues may not differ from those linked to other AI systems, new ones have emerged, for example, in the reasoning behind the Italian data protection authority's decision to suspend the operations of OpenAI's ChatGPT. We provide recommendations regarding what this report considers to be missing data protection provisions, such as reemphasising certain GDPR obligations, including establishing a legal basis for data processing in the context of GAI systems (8).[5]

Eighthly, provisions on the **research exemption** require clarification and changes to support relevant actors (especially the research community) in effectively evaluating and assessing GPAI systems (9).

Finally, the report analyses the AI ACT's **standardisation mechanisms**. It proposes alternative regulatory mechanisms and solutions to promote the inclusion of civil society and affected communities and prevent a potential democratic deficit often associated with standardisation (10).

At the time of writing, the final version of the AI Act still needed to be published and was yet to be developed in a series of technical meetings. Therefore, where reference is made to concrete legal provisions, we refer to the text from the Spanish Presidency from 27 November 2023, respectively, to the earlier texts from the European Council, the Parliament and the European Commission (which were the texts available at the time of the workshop). When referring to concrete legal provisions, we will make explicit which version we refer to.

[4] Veale, Matus and Gorwa (n 2) 15; 'Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer' (*Time*, 18 January 2023) <https://time.com/6247678/openai-chatgpt-kenya-workers/> accessed 7 November 2023.
[5] Ashley Belanger, 'OpenAI Gives in to Italy's Data Privacy Demands, Ending ChatGPT Ban' (*Ars Technica*, 5 January 2023) <https://arstechnica.com/tech-policy/2023/05/openai-gives-in-to-italys-data-privacy-demands-ending-chatgpt-ban/> accessed 7 November 2023.

2. Instruments of Democratic Oversight and Modes of Regulation

The AI Act is intended to follow a proportionate risk-based approach. Also, the political agreement seemed to have landed on systematic risk monitoring obligations, at least for the larger, so-called 'high impact' models. For those high-impact models, the performance of risk assessments is therefore essential to ensure compliance with the AI Act, but even for foundational models that may not fall under the high-risk category, (voluntary) risk assessments can be an important element towards responsible development and deployment of GPAI. Risk assessments (often also called impact assessment) serve at least three goals: influencing design processes to account for individual and societal harms, creating knowledge and transparency to enable learning, and holding the relevant actors accountable.[6] Yet, whether or to what extent a risk-based approach can achieve its normative goals relies upon a risk assessment framework that:

- addresses and actively engages the relevant actors and stakeholders in a substantive manner, including representatives of affected people, especially members of marginalised communities;
- is effective in the sense that it succeeds in creating the right incentives to develop, design, and deploy AI systems in a way that they respect fundamental rights and public values and that it has teeth in case of non-compliance;[7]
- is well-timed in the sense that it can inform the design process to address risks before they materialise and, in the case of risks that materialise only later, recognise and mitigate those risks;
- creates the necessary level of actionable transparency, knowledge gains and transfers.
- creates effective procedural and substantive mechanisms of public oversight, inclusivity, and accountability;
- draws on interdisciplinary expertise from academia, including computational sciences, social sciences and humanities, and civil society, enabling assessors to understand the technology, internal routines, dynamics, motivations, and societal impacts. Likewise, the state-of-the-art should be delineated from an interdisciplinary perspective;

Considering the uncertainties in predicting the impact of the (large scale) deployment of foundation models, risk assessment and mitigation need to be continuous. Similarly, it is important to monitor for individual instances of potential harms and systemic risks for society.

In identifying and mitigating all the risks that may result from developing or deploying a foundation model, the responsibility to perform a risk assessment cannot exclusively fall on the provider, as the deployer decides if and how to use a foundation model in a concrete context. To draw an analogy, one could think of foundation models as a "raw" material, like steel. Suppose a producer produces and sells steel to a bridge builder, and the bridge fails. In that case, there is a need to determine if the failure resulted from how the bridge builder

[6] Andrew D Selbst, 'An Institutional View of Algorithmic Impact Assessments' (2021) 35 Harvard Journal of Law & Technology 117; Alessandro Mantelero, 'Human Rights Impact Assessment and AI' in Alessandro Mantelero (ed), *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI* (TMC Asser Press 2022) <https://doi.org/10.1007/978-94-6265-531-7_2> accessed 15 January 2024.

[7] Selbst (n 6).

used the material (e.g. faulty engineering in a particular context) or if the material itself failed because it was poorly manufactured (at the model level, so to speak). This does not mean that the provider can shift this responsibility by reference to the deployer. Instead, we need due diligence obligations along the value chain:

- developers of the foundation model (the provider) are in a position to assess and mitigate context-free risks that are the result of the way the model has been originally trained, including any risks to the rights of third parties (such as data protection, copyright, and non-discrimination) that are the result of the selection of training data, or risks to the environment, internal and external workers' rights or rights of third parties, including data subjects, copyright holders and affected communities as a result of training and developing the original model.

- deployers can assess and mitigate risks resulting from the re-training, implementation or integration into new services to the health, safety, security, fundamental rights, the environment, democracy or the rule of law protections of users. At the same time, the AI Act will need to specify when deployers turn into developers themselves.

- legislators are in the position to include the relevant cooperation and transparency obligations as well as mechanisms to reverse the burden of proof while considering proportionality requirements to take into account the size, scale and, thus, potential impact of a model.[8] This means that providers and deployers would need to prove that they have taken all means available to them to assess and mitigate reasonably foreseeable risks.

In an area of fast-paced technological development and high-value chain complexity,[9] the difficulty of predicting "reasonably foreseeable risks" in a way that considers the diverse impacts AI systems can have on society at large cannot be overstated[10] and may require new methods of participatory foresight, for example scenario-based methods.[11] In this context, knowledge-sharing mechanisms should be implemented to build expertise concerning the performance of impact assessments and the external evaluation thereof. This expertise should not remain the privilege of a select group of corporate or institutional actors, but be

[8] See recommendations below.

[9] Alex C Engler and Andrea Renda, 'Reconciling the AI Value Chain with the EU's Artificial Intelligence Act' (CEPS 2023) <https://www.ceps.eu/ceps-publications/reconciling-the-ai-value-chain-with-the-eus-artificial-intelligence-act/> accessed 28 November 2023; Ian Brown, 'Allocating Accountability in AI Supply Chains: A UK-Centred Regulatory Perspective.' (Ada Lovelace Institute 2023) <https://www.adalovelaceinstitute.org/resource/ai-supply-chains/> accessed 5 September 2023; Sabrina Küspert, Nicolas Moës and Connor Dunlop, 'The Value Chain of General-Purpose AI' (Ada Lovelace Institute 2023) <https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/> accessed 28 November 2023.

[10] Rishi Bommasani and others, 'On the Opportunities and Risks of Foundation Models' (arXiv, 12 July 2022) <http://arxiv.org/abs/2108.07258> accessed 28 November 2023.

[11] Katharina Messmer and Martin Degeling, 'Auditing Recommender Systems' (Stiftung Neue Verantwortung 2023); Natali Helberger, 'FutureNewsCorp, or How the AI Act Changed the Future of News' (2024) 52 Computer Law & Security Review 105915.

widely disseminated to the benefit citizens, civil society organisations, academia and other stakeholders.

- A European Commission implementing Act or the AI Office should issue more concrete guidance on the methods, internal procedures, and documentation of risk assessments, as well as guidelines for the constructive involvement of independent experts and representatives of affected communities.

- Independent experts should play an important role in identifying and assessing risks for fundamental rights and public values, the environment, workers' rights, etc. In other words, the first step towards an effective risk assessment is identifying which values may be affected in the first place and doing so in a way that accounts for the fact that GPAI will permeate all aspects of society and affect different users or groups in society differently. As these investigations necessitate both technical and societal expertise, efforts to evaluate and examine AI systems should be interdisciplinary. Moreover, for independent experts to be able to play their role, the necessary level of transparency, access to data and reporting obligations need to be legally mandated. Right now, the AI Act does not include an exemption for access to data research comparable to Art. 40 of the DSA.

- To guarantee accountability and effectiveness, an independent external auditing obligation is needed.

- To build expertise, transparency obligations should be strengthened.

3. Accountability along the Value Chain

Under the current versions of the AI Act, the provider or developer of a GPAI is designated as the main actor responsible for ensuring compliance with the AI Act's core obligations. Because the future risks associated with the use of foundation models, and their actual manifestation, remain uncertain, their providers moreover enjoy a great margin of discretion in decisions that are inherently politically charged.[12] For example, they are the ones to initially specify what the relevant design choices will be and what potential data gaps and shortcomings exist. In addition, in the case of foundation models, providers often hold the power to specify and impose their contractual terms on downstream deployers, further expanding their control over how systemic societal and fundamental rights risks will be addressed.

In practice, however, the development and deployment of foundation models involve a **complex value chain of actors**.[13] In many cases, the responsibility for the safety and

[12] Helberger and Diakopoulos (n 3). Nathalie A Smuha and others, 'How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act' (5 August 2021) <https://papers.ssrn.com/abstract=3899991> accessed 21 July 2023.
[13] Ian Brown, 'Expert Explainer: Allocating Accountability in AI Supply Chains' <https://www.adalovelaceinstitute.org/resource/ai-supply-chains/> accessed 1 November 2023.

compliance with the obligations of the AI Act cannot be allocated solely to either the provider (which makes critical decisions about the training data, parameters, training of the model, labour conditions, ecological footprint etc.) or the deployer (which determines how the model is used in a concrete context).[14] While economic and business interests may incentivise companies to keep risks and responsibilities low and indemnify contractually against risks caused by the use of their AI systems through third parties, there can be a societal interest in ensuring a system of cooperative responsibility among operators in the sense of having shared responsibility and a division of duties, while clearly defining each parties' respective obligations.[15] Concretely, this means:

- **Shared responsibility:** Making sure GPAI models are developed and used in a responsible way and in respect of fundamental rights and public values as a shared responsibility of the various actors along the value chain.
- **Existing power dynamics** should be considered, and the instructions or terms of use should not be abused to exclude all responsibility unilaterally, for example, for all high-risk uses (as proposed, however, by the Council's original common position in Article 4c (1) and (2)). Seeing the challenges in identifying any (systemic) risks from the way potentially millions of users use general GPAI systems, there is a real risk that Terms of Use will be used to govern downstream relationships and exclude liability for developers. At the same time, the AI Act will, in all likelihood, not guide deployers on what responsible use of GPAI means.
- **Contractual fairness:** More generally, Terms of Use are a means to impose unilaterally contractual conditions that can be unfair, contrary to good or desirable business practices, or exploit differences in negotiation power, also in relation to foundation models.[16] Therefore, it is important to include a legal mandate to scrutinize the fairness and adequacy of contracts for foundation models, while including a possibility to account for information and power asymmetries, respectively significant differences in size when defining what is fair.
- **Cooperation obligations:** Adding an obligation of mutual assistance and cooperation along the lines of Article 4b of the Council's Common Position, but not limited to the use of foundation models in high-risk areas only.
- **Obligations for deployers:**

  a. To close the accountability gap, it is necessary to also include an obligation for deployers to share with providers and other downstream users, including end-users, where necessary, information needed for compliance with the AI Act (e.g. to identify reasonably foreseeable risks or problems around the accuracy of a model). Additionally, there should be an obligation for developers and providers to create an easy and accessible possibility for (voluntary) incident reporting.
  b. In addition, deployers of foundation models, as the ones that determine a particular context of use, should also be obliged to perform a risk assessment (see the previous

[14] Philipp Hacker, Andreas Engel and Marco Mauer, 'Regulating ChatGPT and Other Large Generative AI Models' (arXiv, 10 February 2023) <http://arxiv.org/abs/2302.02337> accessed 6 March 2023; Helberger and Diakopoulos (n 3).
[15] Natali Helberger, Jos Pierson and Thomas Poell, 'Governing Online Platforms: From Contested to Cooperative Responsibility' (2018) 34 Information Society : An International Journal 1.
[16] Brown (n 13) 39–40.

section on risk assessments) as well as a fundamental rights impact assessment. The possibility to differentiate between larger and smaller deployers must be provided.

- **Open source:** set proportional requirements for "foundation models," recognizing and distinctly treating different uses and development modalities, including open-source approaches.[17]

4. Transparency and Rights of Affected Persons

4.1. A Rare Right for Persons Affected by AI Systems – Art. 52 of the AI ACT

Upon its introduction, the proposed AI Act was meant to ensure the development and deployment of artificial intelligence in compliance with fundamental rights obligations. Where sectors and applications are labelled as high-risk, the reason for doing so, can often be traced back to the system's impact on fundamental rights. A prime example thereof can be found in recital 35 of the Commission's original proposal: "When improperly designed and used, such systems may violate the right to education [Art. 14 CFREU] and training as well as the right not to be discriminated against and perpetuate historical patterns of discrimination [Art. 20 and 21 CFREU]."[18]

Yet, upon analysing the text's three core iterations (Commission, Council and European Parliament), one soon finds a striking omission: the AI Act provides little (in the case of the European Parliament Draft) to no (Commission and Council text) exercisable rights to affected individuals and social groups to protect their fundamental rights and interests. What affected persons are offered is Art. 52: a right toward artificial awareness. A compromise text that was agreed on in December 2023 introduces additional specific transparency obligations for providers and deployers of foundational models. Providers "shall ensure the outputs of the model or system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall also ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account specificities and limitations of different types of content, costs of implementation and the generally acknowledged state-of-the-art, as may be reflected in relevant technical standards. And regarding deployers, the text distinguishes between systems that generate or manipulate image, audio or video content, or text. Regarding the former, deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated. Deployers of an AI system that generates or manipulates text, shall face the same obligation in cases of publications liable to significantly influence the public opinion, unless the text has undergone human review and is subject to editorial control. The information must be provided in a clear and distinguishable manner at the latest at the time of the first interaction or exposure, and

[17] Open Future, 'Supporting Open Source and Open Science in the EU AI Act' <https://openfuture.eu/wp-content/uploads/2023/07/230725supporting_OS_in_the_AIAct.pdf> accessed 1 December 2023.
[18] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts, COM/2021/206 final, Brussels, 21 April 2021.

shall respect the applicable accessibility requirements. Finally, it is up to the European Commission, together with the AI Office, to issue further guidance on the implementation of the transparency obligations (Art. 82 b AI ACT).

Transparency is a necessary but insufficient condition for protecting democratic values and fundamental rights, freedoms and interests. Information holds both intrinsic and instrumental value. Yet, we should extend information obligations to cover the former only. If citizens and those who represent their interests want to tap into the instrumental potential of information, their ability to do so depends upon the available tools. The instrumentalization of information constitutes (at least) three components: reflexivity, observability and actionability.[19] Whereas the former is directed toward the users of AI themselves, the latter targets third parties.

- **Transparency mandates active reflection:** Efforts to make foundation models more transparent should be based on an active reflection process on behalf of their creators, providers and deployers concerning the need for and purposes for which systems will be used, making explicit the choices they make and the assumptions they build on. Risk assessments are one way to effectuate this reflection process, and a robust framework for their performance is therefore essential if the AI Act is to realise its ambitions.

Information should enable citizens to observe the functioning of socio-technical systems within the wider social and cultural context in which these systems are embedded. Observability, in turn, allows citizens to critically interrogate and scrutinise the systems they interact with, which leads to actionability. Information is actionable if it enables citizens to exercise their rights and interests against AI in a manner that is heard and recognised. This too can be understood in a twofold manner. In order to exercise their rights, however, citizens should also be told what their rights would be and *how* they can exercise them.

- **Transparency should enable observability:** Transparency obligations should enable regulators, civil society organisations, and individuals to gauge, scrutinise and contest AI systems' short- and long-term impact on social and democratic values and the fundamental rights and interests of individuals and (social) groups.[20] Transparency requirements should facilitate the observation of AI systems in isolation and as part of the larger social and socio-technical structures with which they interact and are embedded.

- **Transparency should be actionable but is only the first step:** Transparency is a precondition for successfully exercising rights and detecting and correcting (digital) injustice or other wrongdoings. At the same time, individuals, academia, and civil

[19] See also: Paddy Leerssen, 'Seeing What Others Are Seeing: Studies in the Regulation of Transparency for Social Media Recommender Systems.' (2023).

[20] Bernhard Rieder and Jeanette Hofmann, 'Towards Platform Observability' (2020) 9 Internet Policy Review <https://policyreview.info/articles/analysis/towards-platform-observability> accessed 19 July 2023; Leerssen (n 19).

society organisations might differ in their needs and capacity to parse information.[21] Hence, transparency must be adapted to its target audience. Having said so, transparency is only the first step to accountability.[22]

For affected persons, the European Parliament's proposal to incorporate into the AI Act a new right to lodge a complaint and a right to an explanation appears the most relevant.[23] Such a right to an explanation would offer any affected person subject to a decision taken by a deployer based on the output from a high-risk AI system which produces legal effects or similarly significantly affects them the right to request a clear and meaningful explanation from the deployer. This explanation contains information on the role of the AI system within the decision-making procedure, the main parameters of the decision taken and the related input data. Moreover, this explanation should be interpreted in light of Art.13.1 AI Act. For instance, users of AI systems should receive (technical) means to render the system's functioning, the output it produces, and the data processed, interpretable and explainable.

4.2. Remaining Challenges

While citizens do find protection in both primary and secondary law (e.g. the General Data Protection Regulation, the Unfair Commercial Practices, the EU Equality Acquis, etc.), the AI Act wrongfully assumes that existing citizen rights are sufficiently robust to protect the fundamental rights, freedoms and interests of the people affected by the development and deployment of AI systems on the one hand and that these rights can be exercised collectively on the other hand. Yet, in recent years, the digital fitness of various pieces of key EU legislation has been questioned.[24] The AI Act's inclusion of a proper right to an explanation would be a step in the right direction. To be actionable, this right would need to be complemented by rights through which affected persons can contest their subjugation to, and outputs produced by, AI-systems. **Second,** it is far from clear whether collective action can be exercised against harm originating from AI used in the public sector.

We therefore believe that current transparency requirements can be further enhanced. We moreover encourage the EU legislator to continue their efforts to investigate the digital fitness of existing legal instruments, and where needed, propose regulatory changes as to

[21] Nicholas Diakopoulos, 'Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism' (2020) 8 Digital Journalism 945.

[22] Berendt, B. (2022). The AI Act Proposal: Towards the next transparency fallacy? Why AI regulation should be based on principles based on how algorithmic discrimination works. In BMUV & F. Rostalski (Eds.), Künstliche Intelligenz - Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa? (pp. 31-52). Tübingen, Germany: Mohr Siebeck, https://www.mohrsiebeck.com/buch/kuenstliche-intelligenz-9783161612992

[23] European Parliament, Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI, 9 December 2023, https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

[24] Janneke Gerards and Raphaele Xenidis, *Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non Discrimination Law* (Publications Office of the European Union 2021) <https://data.europa.eu/doi/10.2838/544956> accessed 12 August 2021; Natali Helberger and others, 'EU Consumer Protection 2.0: Structural Asymmetries in Digital Consumer Markets' (BEUC 2021); João Pedro Quintais, 'Generative AI, Copyright and the AI Act' (*Kluwer Copyright Blog*, 9 May 2023) <https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/> accessed 27 July 2023.

capture and address the specific risks generative technologies pose to the fundamental rights, freedoms and interests of people affected. Likewise, Article 52 should be complemented with an additional set of exercisable rights to protect citizens against AI-generated harm. These recommendations should be viewed in light of the complexity of the general-purpose AI value-chain [25], and the risk of accountability gaps emerging therein, which highlight the important function information has in facilitating observability and actionability: professional users of general-purpose AI should not be able to hide behind the opacity of complex socio-technical environments. [26]

4.3. Recommendations

Enhanced Transparency for Affected Persons

The information obligations contained within Art. 52 should be further strengthened. Like the Parliament draft, our recommendations focus on both the substance (the what) and modalities (the how) of information provision. Though what mode of information provision is considered appropriate depends upon the context (e.g. vulnerable people), information should be visible in dense policies.

- Information should not only extend to the interaction citizens have with AI, but also cover the envisaged consequences of such interaction, and where possible, the risks thereof, including, but not limited to, manipulation (including the generation of deceitful content and misinformation); bias and discrimination; reduced data protection and privacy; incorrect attribution of content protected by intellectual property rights; and the system's environmental impact.

We understand that it can be challenging for developers and deployers of an AI system to envisage the consequences or adverse impact of their systems, and to provide information on those consequences in a timely, clear, intelligible and visible manner. In this context, however, a mandatory performance of (systemic) risk assessments would actively encourage AI stakeholders to reflect upon the risks their systems could pose to (groups of) natural persons. Users of AI systems should adopt a due diligence approach throughout the various phases of AI development and deployment: they should reflect upon and document their practices. Documentation should remain accessible and scrutable, and in this particular context, help in translating the functioning of AI systems in an understandable manner to affected persons.

- Impact assessments should be made publicly available and easily accessible for affected persons.

- When, as part of their (systemic) risk assessment duties, risks have been identified by users of AI, those who interact with, or are subject to, those systems, should be

[25] See also the section on "AI's distinctive characteristics" in: Brown (n 9).

[26] These issues are analysed in the UK context, which for now still has a similar legal framework to the EU's, in Michael Birtwistle and Davies, 'Regulating AI in the UK: Three Tests for the Government's Plans' <https://www.adalovelaceinstitute.org/blog/regulating-ai-uk-three-tests/> accessed 31 October 2023.

informed on the mitigation strategies that have been implemented, or are still envisaged, to address those risks.

- In addition, persons affected should be informed about the remedies and means that are available to them through which they can oppose, object and/or contest the use of AI-systems.

- In case remedies available include substantive rights, whether granted by the AI Act or other EU instruments, those affected should be informed of the existence thereof in an easily accessible manner upon and during their interaction with these systems, including how these rights can be exercised, either individually or collectively.

Finally, the requirement that information should be offered in a timely, clear, visible and intelligible manner, can be further specified.

- Where possible, information should be contextualised to take into account the recipient's personal situation. Where it is known, or can be anticipated, that an AI-system might interact with persons that are more vulnerable as a result of their age (e.g., children or the elderly) or health (e.g., persons with a visual or auditory impairment), affected persons should have the option to choose between different modalities of transparency befitting their needs. This transparency requirement should not be viewed as an excuse to collect and process additional personal data and special categories of data in particular.
- Different modalities of information provision should be foreseen depending on the (physical) mode of interaction (e.g., text-based chat-bots versus voice assistants).

Finally, data labourers or workers should also be protected under Art. 52 AI Act when they interact with AI systems as part of their activities (see also section 5). This protection comes on top of the recommendations we have formulated regarding the protection of data workers.

Enhanced Contestation Mechanisms for Affected Persons

Additional contestation mechanisms and procedural guarantees should be available to citizens to mitigate and protect them in their exposure to (novel) risks associated with the deployment of GPAI systems, and provide them with the ability to complain, contest and scrutinise the potential for AI systems to interfere with their rights and interests.

- **Right to an explanation:** In case foundation models or generative systems are relied upon to assist or take decisions pertaining to individuals, citizens should have a right to an explanation about that decision, which includes information on the grounds upon which that particular form of assistance or decision was based, the assumptions underlying these operations, and the remedies they have (legal or otherwise) to mitigate/reduce/contest the use of AI and the harm or damages generated.

- **Right to complain:** In addition to their ability to lodge a complaint with a national supervisory authority, (groups of) natural persons affected should also have the opportunity to lodge a complaint about their interaction with AI, and the harms it might have caused them, against any actor in the AI value chain, but at least to the party they interacted with directly as an individual. The AI Act should give people – or the organisations representing them – the opportunity to flag and notify authorities in case they believe they interact with non-compliant AI.

- **Right to effective legal proceedings:** As suggested by the European Parliament, and without prejudice to any available administrative or non-judicial remedy, each natural person should have the right to an effective judicial remedy where they consider that their rights and freedoms have been infringed as a result of non-compliance with this Regulation. They should be guaranteed the right to bring proceedings before the courts of the Member State in which the operator has an establishment. It should also be possible to bring such proceedings before the courts of the Member State where the natural person has their habitual residence, place of work, or the place where the alleged infringement took place.

- **Right to a Human-in-the-loop:** People should have the right to consult with a Human-in-the-loop concerning the use and deployment of AI systems. In addition, people should be able to consult with a person to seek an explanation, file a complaint or exercise an effective remedy. This right enhances and complements the Parliament's proposal to develop AI systems that humans can control and oversee.

- **Right to an accountable value-chain:** Regarding the exercise of their rights, Art. 28.3 EP draft could be amended to include: All operators falling under this Regulation shall ensure that the Regulation provides effective and complete protection for the natural persons affected by an AI system, including coordinating with all other operators in the supply chain. A natural person affected by an AI system may exercise his or her rights under this Regulation towards each of the operators active in such a supply chain. Given the complexity of the value chain, affected persons should not be (solely) responsible for identifying the shares of responsibility of different operators for a given outcome. Instead, in case of a violation, the burden of proof should be reversed: the operator should demonstrate that they were not the one responsible for a given violation.

- **Right of Voice:** People should have a right not to have their (personal) data used for further training due to their interaction with foundation models and systems built through them.

Enhanced collective ability to contest AI systems:

Lawmakers should promote mechanisms that facilitate this to ensure that affected persons can exercise their rights effectively and collectively, whether through a representative organisation or not.

a. **Enhanced investigative capabilities:** The introduction of a joint investigation provision in the AI Act EP draft is to be commended (Art. 66. a). However, its scope of application is too narrow. Though the DSA/DMA model is good as it centralises investigations, it applies only where 45 million monthly active users are affected. The law should make joint and cross-national investigations possible where smaller but potentially equally harmful foundation models are rolled out. In addition, non-profit and civil society organisations should be able to launch an investigation into the practices of AI actors and flag potentially non-compliant use on their behalf. To do so, however, it is important to increase public transparency obligations (as recommended above).

b. **Right to collective redress:** Though the addition of the AI Act to the Collective Redress Directive is a step in the right direction to increase the collective ability of citizens and civil society organisations, the AI Act should directly and explicitly foresee collective redress or representation options. This would enable non-profits to launch investigations and actions to benefit society in all areas covered by the AI Act, including public contexts, such as law enforcement and welfare, rather than consumer contexts only.

c. **Right to mandate a third party to exercise rights:** A natural person should have the right to require a (not-for-)profit body, organisation or association which has been properly constituted in accordance with the law of a Member State, has statutory objectives which are in the public interest, and is active in the field of the protection of rights and freedoms of natural persons, to exercise the rights of such persons under this Article.

Reversal of burden of proof:
Above, we argue in favour of increased accountability. More specifically, affected persons should be able to exercise their rights against any operator within the AI value chain. Yet, even if affected persons can exercise their rights against anyone, due to the complexity of the digital ecosystem, they would still face difficulties in demonstrating and proving they have been wronged or exposed to undue risk. Similar to the EU non-discrimination law, the AI Act should foresee mechanisms that reverse the burden of proof:

- Upon having identified plausible harm, the onus would be on the operator to prove they did all they reasonably could to avoid reasonably foreseeable negative impact.

- In case an operator successfully demonstrates they did all they reasonably could to avoid the harm in question, this should not result in a situation where those affected can no longer be compensated for the damages they did suffer. Hence, alternative redress and compensation mechanisms should be considered where no direct responsible party can be found.

Enhanced transparency for external accountability purposes (including risk assessments)

Knowledge-sharing mechanisms should turn transparency into an active practice. Third parties, like citizens, academia, and civil society, should be able to scrutinise the underlying assumptions and the choices made before and during the development of the foundation model or (modified) derivatives.

Registered information should be made available in two tiers. In the first order, the information shared should keep its original granularity and complexity to enable scrutiny by regulators, (academic) experts, auditors and civil society. In the second order, this information should be clearly communicated in a manner that is accessible and easy to understand for the broader public. To facilitate this process of knowledge exchange and external accountability, **the following information should be made public by both providers and deployers as part of Annex VIII's Registration duties:**

1. Fundamental Rights and Data Protection Impact Assessments should be made available in full rather than in a summarised version.

2. Information on data and data governance (Art. 10 AI Act) and technical standards utilised (Art. 11 AI Act and Annex IV Technical Documentation).

3. Documentation concerning data governance requirements should be extended to include information on the data production process, such as third parties relied upon during the data production process and instructions provided to staff or third parties concerning the labelling and training of data.[27]

4. Information requirements listed in recitals 60g and 60h of the European Parliament Compromise Text, including, among others, documentation concerning technical and data governance strategies used to mitigate risks and harms concerning the system's performance, predictability, interpretability, corrigibility, safety, cyber security, environmental and fundamental rights impact.

5. Documentation of the periodic monitoring mechanisms that take place or are envisaged.

6. Providing information on the data and data sources should not interfere with the fundamental rights and interests of those natural (and legal) persons whose data are part of the training and collected data.

5. Environmental Impacts

[27] Milagros Miceli and others (2022), 'Documenting Data Production Processes: A Participatory Approach for Data Work'. Proceedings of the ACM on Human-Computer Interaction, Volume 6, Issue CSCW2, Article No.: 510, pp 1–34. https://doi.org/10.1145/3555623; Milagros Miceli and Julian Posada (2022), 'The Data-Production Dispositif'. Proceedings of the ACM on Human-Computer Interaction, Volume 6, Issue CSCW2. Article No.: 460. pp 1–37. https://doi.org/10.1145/3555561.

Even before Open AI's ChatGPT popularised GPAI systems worldwide, researchers have been trying to bring attention to the environmental impacts of training AI systems and the carbon effects linked to big tech's infrastructural aspects of AI deployment.[28] In addition, AI models' water footprint has recently started undergoing increased scrutiny.[29] One of the main issues in diminishing AI's environmental impact is the need to quantify the latter and transparently inform about the results and measures taken to mitigate energy consumption and carbon emissions. The AI Act must not miss the opportunity to bring attention to those topics and to reduce environmental risks linked to the development and deployment of GPAI. Under the compromise proposal, considering the ecological impact is delegated to a large extent to codes of conduct and additional guidance that the EC develops on the implementation of Art. 12 and 28b.

As discussed in the section on **risk assessments,** there is a need to assess and mitigate context-free risks resulting from how the model was originally trained, including those posed to the environment. Similarly, deployers should also assess and mitigate, among others, environmental risks resulting from the re-training, implementation or integration into new services. In other words, considering the ecological should be part of the mandatory and voluntary risk assessments.

Information that should be made public by both providers and deployers as part of Annex VIII's Registration duties should include information requirements listed in recitals 60g and 60h of the original European Parliament Compromise Text, such as (among others) documentation concerning technical and data governance strategies used to mitigate risks and harms concerning the system's environmental impact.

A law that should be considered complementary in the context of the AI Act's provisions on the environment is the **Corporate Sustainability Directive**. The directive aims to ensure that companies operating within the European Union take responsibility for addressing human rights and environmental risks throughout their global value chains. This should include the potential environmental impact (for example, through resource consumption) of training AI models.

6. Workers' Rights

The AI Act proposal demands rights for end users and "affected persons" (such as societal groups represented in AI in discriminatory ways). The European Parliament version rightly extends the stakeholders to be considered to the planetary/ecological environment. However, an important group of people strongly affected and often harmed by AI development are the workers who help build AI systems, for example, by performing data labelling or content

[28] Payal Dhar, 'The Carbon Impact of Artificial Intelligence' (2020) 2 Nature Machine Intelligence 423.
[29] Pengfei Li and others, 'Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models' (arXiv, 6 April 2023) <http://arxiv.org/abs/2304.03271> accessed 20 October 2023.

moderation tasks.[30] These workers' rights must also be protected to achieve the AI Act's goal of safeguarding fundamental rights and freedoms.

As a result, there is a need to refer explicitly to the rights of information workers as part of the risk assessment, as part of appropriate data governance measures and the role of workers in the production of those data sets, and train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the violation of workers' rights (including data workers involved in the development of generative AI systems).

Concerning other relevant legislation regarding workers' rights, **the Platform Work Directive and the AI Act should complement each other**. This directive aims to protect the rights of the EU's gig economy and platform workers. As mentioned previously, workers' rights are relevant for GPAI, as many of them contribute to developing and fine-tuning the models, for example, by labelling data and detecting illegal content or hate speech. In addition, work platforms may use GPAI to automate management processes that directly affect gig workers.

In the context of the AI Act's provisions on workers' rights, the **Corporate Sustainability Directive** could once again be critical. The directive aims to ensure that companies operating within the European Union take responsibility for addressing human rights risks throughout their global value chains. This should include the working conditions of workers in the global south who contribute to the training of GPAI models (in addition to the environmental impacts mentioned in the section on the environment).

Finally, as discussed in the sections on **risk assessments and the environment,** there is a need to assess and mitigate context-free risks resulting from how the model was originally trained, including those posed to workers' rights.

7. Copyright

Copyright-related issues in GPAI are crucial for authors and those who want to benefit from generative AI systems. Firstly, it is important to ensure that the development of AI technology does not unduly harm the rights of creators (authors and performers) and other rights holders, as recognised and protected under the EU copyright acquis. Secondly, those using generative AI need certainty that they comply with relevant copyright rules.

The compromise version of the AI Act agreed on December 6-8, 2023, includes two provisions dealing with using copyrighted material as part of training data. These provisions build on the approach first proposed in the foundation model section of the European Parliament Report. They are contained in a new section 52c on "Obligations of Providers of GPAI Models." Section 1(c) will require all providers of General Purpose AI models to…

[30] Adrienne Williams, Miceli Milagros and Gebru Timnit, 'The Exploited Labor Behind Artificial Intelligence' (13 October 2022) <https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence> accessed 20 October 2023.

> *put in place a policy to respect Union copyright law, in particular, to identify and respect, including through state-of-the-art technologies, the reservations of rights expressed under Article 4(3) of Directive (EU) 2019/790;*

And section 1(d) requires them to

> *draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office;*

The latter provision is a clear improvement over the original Parliament text, as it no longer suggests that model providers need to distinguish between copyright-protected and public-domain training material and then apply different transparency standards to each, which would be unworkable. A proposed recital by the European Parliament[31] clarifies what would constitute a "sufficiently detailed summary". The recital in question makes it clear that the "summary should be comprehensive in its scope instead of technically detailed, for example, by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used". It also emphasises that the template to be provided by the AI Office should "allow the provider to provide the required summary in narrative form".

The obligation to "train, design and develop foundation models to ensure adequate safeguards against the generation of illegal content under EU law, which includes copyright infringing content, in line with the generally-acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression" that was contained in the Parliament's position has not made it into the final compromise. For copyright, this likely means that technological safeguards against the generation of copyright-infringing content by GPAI models will primarily be left to private ordering, namely contractual arrangements between the providers of those models and rights holders with sufficient bargaining power. Such safeguards will likely be in the form of content filtering measures, designed and primarily informed by the commercial interests of model providers and rightsholders rather than freedom of expression concerns.

The final compromise is a reasonable outcome. Instead of introducing new rules outside the copyright acquis to address the issues raised by the use of copyrighted works to train generative AI models, the text refers back to the existing rules for text and data mining in the 2019 Copyright in the Digital Single Market Directive (CDSM) Directive (Directive (EU) 2019/790). Here, the text provides additional clarity on the compliance obligations of providers of generative AI models (Article 52(1)(c)) and addresses the transparency concerns raised by creators, rights holders, academics and civil society organisations (Article 52(1)(d)). The transparency obligation, in particular, could prove instrumental in assessing whether GPAI models were developed from lawfully accessible copyrighted training

[31] At the time of writing the recitals were still being discussed in technical meetings so this language could still change.

materials, a crucial requirement to carry out lawful text-and-data mining under the CDSM Directive.

This approach further underlines the need for more standardisation regarding machine-readable reservations based on Article 4(3) of the CDSM Directive.[32] The reference to "state-of-the-art technologies" in the text of the AI Act points to the need to develop technological frameworks that allow for the exchange of rights reservation or “opt-out” information. This is an area where the EU should take a more active role, in parallel with the implementation of the AI Act, to ensure that rights holders and AI developers can rely on a robust set of standards.

In general, it is welcome to see that despite the high-stakes negotiations, EU lawmakers have agreed on an approach that reinforces the balanced approach to allowing the use of copyrighted works for generative AI training established by the TDM exceptions introduced in the CDSM Directive. By improving transparency of training data and the possibility that rights holders can reserve their rights while at the same time ensuring that AI models can be trained on other legally available information, the EU has created a framework that provides additional legal certainty without having to make a binary choice between the interests of creators on the one hand and AI developers on the other.

8. Data Protection

The recent large-scale release of GPAI systems has led to data protection concerns.[33] A few months ago, following an intervention by the Italian data protection authority, OpenAI was forced to refrain from processing Italian citizens' personal data. While this prohibition has been temporarily ended, an investigation is still being conducted and could result in new restrictive measures in the future. This is, of course, not only an Italian concern. Other European data protection authorities are also looking into this topic. Issues such as transparency measures to inform about personal data collection processes by general GPAI systems or the validity of a legal basis to process people’s personal data need urgent attention.

- **Reasonable expectations**: in terms of data protection, the scraping of people’s data, its use for training and, subsequently, this data being revealed in a new context is not what European citizens would reasonably expect. It would constitute a violation of Art. 8 of the Charter and GDPR’s obligation for data processing to be lawful, fair and transparent;
- **Legal basis:** certain issues that the AI Act would like to tackle could be solved by properly enforcing other laws, including the GDPR, non-discrimination law and consumer protection law. In that light, the regulation should re-emphasise relevant GDPR requirements, including the necessity of a legal basis to process personal data;

[32] Article 4(3) CDSM Directive reads: “The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.” See also supporting recital 18 CDSM Directive.
[33] Brown (n 13) 12–16.

- **Transparency**: developers of GPAI models should be transparent about personal data used to train their systems. Risks related to generating biased data should be explained;
- **Diversity of data subjects:** transparency measures should be adapted to the needs of user groups with diverse backgrounds and knowledge interests;

9. Research Exemption

Researchers have an important role as innovators and critical observers. On the other hand, they have responsibilities not to harm or abuse their position to circumvent the rules. This is why it is so important to include researchers in the act and provide for a solid research exemption. The AI Act calls on the Member States to support and promote research and development of AI solutions and in this way also support socially and environmentally beneficial outcomes. The strong focus on the important societal role of research is to be welcomed. To be able to play that role, researchers, however, need the active support of national governments and the European Union to build capabilities, attract and keep talent from technical sciences, and social sciences and the humanities, train new generations of interdisciplinary scholars but also find workable and meaningful ways for academic researches to share their findings while being respectful of how academia works.

There are carves out for research to allow experimentation, promote innovation, prevent harm and enable testing of the AI systems before their potential release to the world. To benefit from the research exemption, research organisations (the nature of which should be clearly stated) should be able to rely on, and also comply with a set of requirements that need to be further defined.

- **Exemption before use:** it should be noted that research must be able also involve developing or tweaking a foundation model after it has been released,
- **Ethical and professional standards:** to ensure that the exemption complies with ethical norms, a research exemption could be complemented by a provision similar to the language adopted in the Council version of the AI Act: 'Under all circumstances, any research and development activity should be carried out in accordance with recognised ethical and professional standards for scientific research'.
- **Beneficiaries of the research exemption:** in general, there is a need to explain whether the exemption would benefit only certain types of organisations. For example, in copyright law, text and data mining (TDM) exceptions for research purposes are only available for 'research organisations', and 'cultural heritage institutions' and similar institutions;
- **Taking into account various risk levels when testing:** it should be made clear through appropriate wording that the exemption needs to be interpreted in a way that would lead to minimum harm. For this reason, research and testing should take into account the risk category of the AI system under development. The level of freedom provided to those conducting tests when developing AI systems should depend on the level of risk associated with the latter (high-risk, low-risk, etc.). The delegated regulation should define contextually adequate minimum technical and security measures to benefit from the research exemption. In case of doubt whether the testing can be

conducted in compliance with relevant norms, it should take place in a regulatory sandbox to ensure strict supervision (in line with the provisions on the establishment of regulatory sandboxes);

- **Open source:** it should be explained to what extent researchers that build or tweak an open source model would fall under a research exemption[34];
- **Prioritising delegated regulation:** exhaustive delegated regulation would be an effective measure due to its enforceability (as compared to, for example, voluntary codes of conduct);
- **Access to data and models:** the AI Act misses a provision of research access for academic and civil society researchers, similar to Art. 40 DSA. For researchers to be able to play their role as critical observers and testers, access rights are quintessential. This is especially important given the increasingly unequal level playing field between public interest research and commercial research in this field.

10. Standardisation

Governance of complex systems like foundation models / general purpose AI consists of a combination of approaches, from the AI Act itself (and beyond) to its implementation acts, common specifications, standardisation, codes of conduct, terms of use, and enforcement. Having said this, strong and enforceable laws are the cornerstone of effective protection of people's rights.

Recently proposed self-regulatory initiatives like a *code of conduct on artificial intelligence* or an *AI Pact* may have an effect in shaping current debates about what society expects from companies developing and deploying such systems. Still, they can be no alternative to setting rules on foundation models / GPAI in the framework of democratic decision-making, as such voluntary codes have consistently failed to protect people's rights effectively.[35]

Likewise, there is a well-known and long-standing problem with standardisation, especially when it is supposed to prevent risks to human rights as part of a highly technical product-safety approach.[36] Moreover, ensuring that civil society and affected communities are adequately represented in the process has proven difficult.[37] The industry is strongly represented in standard-setting, adding to the potential democracy deficit.

To address these challenges, the AI Act, with regard to foundation models / general purpose AI,

[34] Open Future (n 17).

[35] 'AI Pact | Shaping Europe's Digital Future' <https://digital-strategy.ec.europa.eu/en/policies/ai-pact> accessed 30 November 2023; 'Hiroshima Process International Code of Conduct for Advanced AI Systems | Shaping Europe's Digital Future' (30 October 2023) <https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems> accessed 30 November 2023.

[36] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach' (2021) 22(4) Computer Law Review International 97.

[37] ibid 105.

- should include a provision giving the EU Commission the power to adopt a delegated act, specifying the methodology for risk and impact assessments for both providers and deployers. This methodology needs to be developed with the involvement of experts, civil society and representatives of affected groups.
- can, however, for certain technical issues, define technical specifications as common specifications in case technical specifications are necessary (similar to the ones mentioned in Art. 41 of the AI Act proposal) in a process that ensures a balanced representation of interests and effective participation of all relevant stakeholders. These issues should not be addressed by technical standards and norms (in the sense of Art. 40 of the AI Act proposal);
- should, for the common specifications, aim for a level of abstraction that is specific enough to be helpful for AI developers and at the same time abstract enough to remain relevant for five to ten years in this fast-developing field;
- should not allow developers of general purpose AI to avoid responsibility in case they "explicitly excluded all high-risk uses in the instructions of use or information accompanying the general purpose AI system" (the foundation model) and have no "sufficient reasons to consider that the system may be misused", but instead include mechanisms of regulatory scrutiny regarding the fairness, quality and adequacy of contractual terms and instructions;
- should regard the systemic risk monitoring approach in Art. Thirty-four of the Digital Services Act (DSA) as an inspiration. Under the DSA, Very Large Online Platforms and Very Large Search Engines are obligated to monitor their algorithmic systems regularly for any actual and foreseeable negative effects on fundamental rights and societal processes, including such that arise from the development, design, and deployment of GPAI. It is conceivable that a comparable obligation to monitor for and mitigate systemic risks regularly should also apply to the providers of general purpose AI.