Policy briefing

October 2023

An EU AI Act that works for people and society

Five areas of focus for the trilogues

This September 2023 briefing updates the Ada Lovelace Institute's recommendations on the EU Artificial Intelligence Act (AI Act) as the trilogues get underway. This mirrors the development of AI technologies, which have progressed at a rapid rate since the AI Act was proposed in April 2021, and particularly over the past year.

Advances in development have led to significant jumps in capabilities (for example, AI 'godfather' Yoshua Bengio told the US Senate: 'I and many others have been surprised by the giant leap realised by systems like ChatGPT').[1] Wide-scale deployment has seen rapid uptake and, at times, disruption of existing sectors and markets (for example, Open AI's release of ChatGPT and its impact on education).[2]

1 Gerrit De Vynck, 'AI Leaders Warn Congress That AI Could Be Used to Create Bioweapons' Washington Post (25 July 2023) https://www.washingtonpost.com/technology/2023/07/25/ai-bengio-anthropic-senate-hearing/ accessed 30 August 2023

2 Nicole Serena Silver, 'The Future Of Education - Disruption Caused By AI And ChatGPT: Artificial Intelligence Series 3/5' (Forbes) https://www.forbes.com/sites/nicolesilver/2023/06/05/the-future-of-educationdisruption-caused-by-ai-and-chatgpt-artificial-intelligence-series-3-of-5/ accessed 30 August 2023

Ada Lovelace Institute

For more information about the Ada Lovelace Institute and our work on the EU AI Act, contact Connor Dunlop: cdunlop@adalovelaceinstitute.org

Experts expect this progress to continue at 'breakneck speed for at least the next few years'.[3] To support policymakers, we have therefore updated our position based on developments in the AI space, our latest research and expert convenings.

Based in London and Brussels, the Ada Lovelace Institute (Ada) is an independent research institute with a mission to make data and AI work for people and society. We maintain an international outlook that recognises the importance of the EU's developing regulatory proposals, both within the EU and globally.

Our work brings together evidence-based research with expert convenings to influence policy and practice – with the aim of ensuring that that AI is developed and deployed in a trustworthy manner, that AI risks are mitigated, and that accountability can be clearly assigned when things go wrong.

The recommendations in this briefing are informed by our research. On the AI Act specifically, in early 2022 we published an expert opinion paper by Professor Lilian Edwards and our 18 recommendations for strengthening the AI Act.

In 2023, we have published reports and briefings on foundation models; general purpose AI and their release strategies; allocating accountability in the AI value chain; risk assessment and mitigation across the AI lifecycle; AI standards and civil society participation; and regulatory functions for monitoring AI development. More information and links are provided at the end of this briefing.

3 '4 Charts That Show Why AI Progress Is Unlikely to Slow Down' (Time, 2 August 2023) https://time.com/6300942/ai-progress-charts/ accessed 1 September 2023.

Summary of recommendations

To strengthen the AI Act we recommend that negotiators focus on five areas:

1. Centralised regulatory capacity to ensure an effective AI governance framework

- Ada supports the European Parliament's proposal for a central AI Office and agrees broadly with how it should function.
- Ada supports the European Parliament's proposal to have the AI Office conduct monitoring and foresight activities.
- Ada supports the European Parliament's proposal to set up permanent sub-groups of AI developers and other relevant stakeholders in the AI Office Advisory Forum, to consider the governance of foundation models and R&D.
- Ada supports the European Parliament's proposal to have the AI Office coordinate cases involving more than one member state, or suspected widespread infringements by high-risk or foundation model providers, including via onsite and remote inspections.
- Ada supports the European Parliament's proposal for the AI Office to be empowered to issue (binding) guidance and analysis on emerging issues, to contribute to standardisation processes and benchmarking, as well as to work on codes of conduct.

2. Fostering responsible development and distribution of foundation models

- The European Parliament's proposed rules for foundation models (Article 28b) should be implemented in full and strengthened further (as detailed in the following bullet points).
- Mandatory disclosure requirements for developers of foundation models operating in the EU should be introduced, including notification to the AI Office or national regulators when beginning large-scale training runs, coupled with disclosure of compute and capabilities evaluations.

- The AI Act should introduce pre-market third-party conformity assessment for foundation models, ensuring the burden of proof is on developers to prove safety and compliance before widescale distribution.
- Foundation models should be rolled out in a staged manner, and provided via 'structured access' – at least until regulators have more visibility over how these models interact with people and in complex environments.
- Developers of foundation models provided via an application programming interface (API) should offer a complaint mechanism for users to swiftly report any serious risks that have been identified. The provider should take corrective action when possible and notify the relevant supervisory authority when not possible.
- Foundation model providers and downstream application providers should label and make clear the underlying foundation model to end users and affected persons.
- Foundation models should be regulated regardless of distribution channel, meaning open-source exemptions should not be applied to this specific type of AI.
- The AI Office should be empowered to add additional obligations for foundation models deemed uniquely capable or strategically significant, or to issue partial exemptions for models deemed to be of public benefit (for example, collaborative development of free open source).
- The legal loophole for general-purpose AI (often referred to as 'GPAI' or used interchangeably with foundation models) in the Council of the EU's general approach (Article 4c) should not be included in the final AI Act, so that developers cannot relinquish responsibility using a standard legal disclaimer.

3. Mitigating risk throughout the AI system lifecycle via an ecosystem of inspection

- Developers and deployers should offer vetted researcher access to carry out external inspections – such as red-teaming – to test AI models for vulnerabilities and risks to health, safety or fundamental rights, as included in the EU Digital Services Act.[4]

- The AI Act should include 'safe harbour' provisions for industry and researchers, designed to reasonably assure that entities participating in good faith auditing exercises on a good faith basis are not subjected to undue liability risk or retaliation.

- An EU Benchmarking Institute should be set up to coordinate benchmarking and national metrology authorities[5] – and other relevant stakeholders – on creating benchmarks and thresholds for measuring safe development and deployment of AI models.

4. Maintaining a risk-based approach and future-proofed regulation

- New filters should not be introduced for high-risk categorisation. The European Commission's proposal for high-risk categorisation should be maintained.

- A clear mechanism for updating the high-risk list of categories in Annex III should be included, as well as retaining the ability to add sub-categories below those existing categories. Affected persons should be empowered to flag any systems that they believe should be added to this list.

- Any changes to the high-risk list should be based on clear, judicially reviewable criteria, and should consider systemic and environmental risk.

---

4 'The Digital Services Act Package | Shaping Europe's Digital Future' (31 July 2023) https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package accessed 5 September 2023.

5 Metrology is the science of measurement and is used across most domains of research and innovation in EU industrial policy (and beyond).

5. Protection and representation for affected persons

- The AI Act should include a definition of 'affected persons' and rename 'users' to 'deployers' to more accurately reflect the AI lifecycle.
- Include an obligation for a pre-deployment fundamental rights impact assessment (FRIA), as proposed by the European Parliament.
- The Act should implement a comprehensive remedies framework for affected persons based on effective avenues for redress, including a right to lodge a complaint with a supervisory authority, judicial remedy and an explanation of individual decision-making – as proposed in Chapter 3a of the Parliament's text.
- Enhance multi-stakeholder participation in standards development processes and enhance transparency over these processes.
- Include a standing panel of representative users or a citizens' assembly as a permanent sub-group of the Office's Advisory Forum.

See detailed recommendations below, including suggestions for how to reflect these in the final text of the AI Act.

1. Centralised regulatory capacity to ensure an effective AI governance framework

To ensure that the EU's regulatory ecosystem has the necessary capabilities to implement the AI Act, a central EU regulator will be best placed to offer 'effectiveness, efficiency, coherence and legitimacy'.[6] To make this a reality, however, it will also have to be adequately resourced.

Resourcing

AI technologies can impact a wide range of sectors and aspects of society, and will increasingly form part of the EU's digital infrastructure. The EU central functions should therefore be funded as much as, if not more than, other domains where safety and public trust are paramount and where underlying technologies form important parts of national infrastructure – such as civil nuclear, civil aviation, medicines, road and rail.[7]

As in already in place in some of these sectors (for example, 80% of the European Medicines Agency's funding comes from the market entities it regulates)[8] a mandatory fee could be levied on developers over a certain threshold (for example, expenditure on training runs, or compute). For more detail see the section below on 'Considering a tiered' approach for foundation models. This fee could be used to build out an AI Office with relevant expertise to adequately fulfil its functions. We therefore encourage policymakers to recognise the need for a strong AI Office for effective enforcement of the AI Act, and to consider comparable models for providing requisite funding.

AI Office functions

The AI Office should lead on key cross-cutting activities, such as enforcement cases involving more than one member state, and monitoring and foresight. This would allow member states to benefit from consistency and uniform application across the EU, and for coordination and shared learning across member states. The central function could also act as a point of expertise and specialisation, and the point of contact for establishing quick feedback loops between policymakers and industry (vital in the ever-shifting AI landscape).

To strengthen cross-cutting activities, the European Parliament's suggestion for the AI Office to coordinate joint investigations based on suspected widespread infringement is also welcome and should be included. This would address concerns voiced around state-level enforcement capacity

6 Nicolas Moës, Felicity Reddel, and Samuel Curtis, 'Giving Agency to the AI Act' (The Future Society, 2023) https://thefuturesociety.org/wp-content/uploads/2023/04/giving-agency-to-the-ai-act.pdf

7 In the UK, the Civil Aviation Authority has a revenue of £140m and staff of over 1000, and the Office for Nuclear Regulation around £90m with around 700 staff. An EU-level agency for AI could reasonably be expected to be funded well beyond this given the vastly larger size of the EU market.

8 EMA, "Funding," European Medicines Agency, Sep. 17, 2018. https://www.ema.europa.eu/en/about-us/how-we-work/governance-documents/funding (accessed Aug. 10, 2023).

('Of the five most populated countries in the EU, only Spain, with a new regulatory AI sandbox and AI regulatory agency, seems to be well prepared')[9] and reduce the risk of divergent levels of protection as seen at times under the GDPR.

There should be strong evidence-gathering mechanisms to address the information asymmetries between AI developers and regulators. This could include a 'supervisory examination' system, in which regulatory authorities would assess and oversee the activities of regulated entities, to ensure their compliance with the AI Act and corresponding standards.[10] We support the European Parliament's proposal for unannounced on-site and remote inspections (for high-risk and foundation models), by either national supervisory authorities or the AI Office.

In relation to monitoring and foresight, there is an existing AI monitoring ecosystem, but it has significant gaps. Centralised EU regulators could address these gaps through their unique powers to access direct, continuous and high-value information from companies, and to collaborate across governments (both at an EU and international level) on comparative and complementary foresight efforts.[11]

Proposals should include establishing a means for quick feedback loops via establishing permanent sub-groups of AI developers and other relevant stakeholders, as well as AI incident documentation (including 'near misses' and allowing reporting from those 'on the ground', that is, whistleblowers and affected persons).[12]

These mechanisms are supported by 96% of respondents in a survey of 51 experts (from AI labs, and those working in civil society and academia on frontier AI governance).[13] They have also been shown to increase trust and to promote a culture of transparency and accountability in other sectors.[14] Such interventions will upskill regulators, who should in turn distribute these learnings to other actors in the AI ecosystem, such as small and medium-sized enterprises (SMEs) and startups, civil society and academia.

As a result of monitoring and foresight, stakeholder dialogues and incident documentation, the AI Office will be better placed to act in an agile way and ensure future-proofed governance. It should therefore have powers to issue (binding) guidance and analysis on emerging issues, to contribute to standardisation processes and benchmarking, as well as to work on codes of conduct, as proposed by the European Parliament.

---

9 'Key Enforcement Issues of the AI Act Should Lead EU Trilogue Debate' (Brookings) https://www.brookings.edu/articles/key-enforcement-issues-of-the-ai-act-should-lead-eu-trilogue-debate/ accessed 30 August 2023

10 The Future Society, 'Blueprint for an AI Office' (2023) (forthcoming)

11 Ada Lovelace Institute, Keeping an Eye on AI: Approaches to government monitoring of the AI landscape (2023) https://www.adalovelaceinstitute.org/report/keeping-an-eye-on-ai/ accessed 30 August 2023

12 Kris Shrishak, 'How to Deal with an AI Near-Miss: Look to the Skies' (2023) 79 Bulletin of the Atomic Scientists 166 https://thebulletin.org/premium/2023-05/how-to-deal-with-an-ai-near-miss-look-to-the-skies/

13 Ada Lovelace Institute, Regulating AI in the UK (2023) https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/ accessed 30 August 2023

14 Ibid

Recommendations

- Ada supports the European Parliament's proposal for a central AI Office and agrees broadly with how it should function. However, the text should clarify that it will be funded to an equivalent level as other domains in which safety and public trust are paramount. The AI Office should hold the right to levy a mandatory fee on designated market players to fund this.

- Ada supports the European Parliament's proposal to have the AI Office conduct monitoring and foresight activities. The AI Office should also document and publish serious incidents including near misses, and affected persons and whistleblowers should be empowered to report such incidents.

- Ada supports the European Parliament's proposal to set up permanent sub-groups of AI developers and other relevant stakeholders in the AI Office Advisory Forum, to consider the governance of foundation models and R&D.

- Ada supports the European Parliament's proposal to have the AI Office coordinate cases involving more than one Member State or suspected widespread infringements by high-risk or foundation model providers, including via onsite and remote inspections.

- Ada supports the European Parliament's proposal for the AI Office to be empowered to issue (binding) guidance and analysis on emerging issues, to contribute to standardisation processes and benchmarking, as well as to work on codes of conduct.

2. Fostering responsible development and distribution of foundation models

Foundation models present novel governance challenges for the AI Act as they do not conform to the product-safety paradigm of being released for a specific use or a specific market.

These challenges centre on:

- the complexity of the AI value chain, and the number of actors building on top of foundation models

- the multi-functionality of foundation models (they can be deployed for a variety of uses with divergent risk profiles) and the tendency for capabilities – and therefore risk – to grow as the models are trained with more data and computing resources

- the disparate levels of control of the models, depending on choices made about how they are distributed (API vs open source).[15]

Foundation models are therefore differentiated from other AI systems (such as more narrow use 'generative AI' systems, which are usually designed for a specific purpose),[16] because risk is harder to locate and can emerge across the AI lifecycle.

Problematic behaviours could originate in pre-training data, for example, or new ones could emerge when the model is integrated into complex environments (like a hospital or a school).[17] This is amplified by the fact that foundation models can be built 'on top of' to develop different applications for many purposes, which means errors or issues at the foundation-model level can quickly proliferate among any applications built on top of (or 'fine-tuned') from that foundation model.[18]

Given their unique features, foundation models hold similarities to other novel, complex and partly experimental technologies with potentially severe consequences (such as Class III medical devices in the USA).[19] [20] Regulators should therefore apply similar standards of care and evidentiary burdens for efficacy and safety. For AI regulation, this should entail:

- disclosure of evaluation and testing plans with a regulator, via mandatory disclosure mechanisms
- various evaluations and evidence points across the foundation model value chain, starting with third-party audits at the pre-market stage
- requirements on testing and evaluations throughout the value chain, leaning on an 'ecosystem of inspection'
- post-market monitoring, via transparency mechanisms, user and affected person reporting, regulator inspection and vetted researcher access.[21]

---

15 Sabrina Küspert, Nicolas Moës and Connor Dunlop, 'The value chain of general-purpose AI' (Ada Lovelace Institute, 10 February 2023) https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/ accessed 30 August 2023

16 Ada Lovelace Institute, 'Explainer: What Is a Foundation Model?' (2023) https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/ accessed 30 August 2023

17 Ada Lovelace Institute, AI Assurance? Assessing and mitigating risks across the AI lifecycle (2023) https://www.adalovelaceinstitute.org/report/risks-ai-systems/ accessed 30 August 2023

18 'Companies that have built and optimized their products to work with a certain iteration of OpenAI's models could "100%" see them suddenly glitch and break, says Sasha Luccioni, an AI researcher at startup Hugging Face. When OpenAI fine-tunes its models... products that have been built using very specific prompts, for example, might stop working in the way they did before.' Source: Heikkilä, M. The Algorithm newsletter, MIT Tech Review, 7.24.23

19 Ada Lovelace Institute, FDA-style oversight for foundation models (forthcoming).

20 'Auditing Algorithms: The Existing Landscape, Role of Regulators and Future Outlook' (GOV.UK) https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook accessed 30 August 2023

21 Ada Lovelace Institute (n 16)

Mandatory disclosure requirements

The first step towards responsible development should be to enable regulators to access insights about foundation models ahead of their release, as there is currently a significant information asymmetry between governments and regulators, and those developing and using these models. One mechanism to address this is mandatory disclosure requirements for developers of foundation models operating in the EU – as used in civil aviation and life sciences.[22] [23]

This should start with notification to the AI Office or national regulators when beginning large-scale training runs of new models, coupled with disclosure of compute and capabilities evaluations.[24] Compute and capabilities are high-value information, as these are indicative of AI models that are particularly likely to precipitate risk or harms.

This proposal builds upon the European Parliament's suggestion (Annex VIII, Section C) for foundation model developers to disclose high-value information (for example, data used to train models, results from in-house audits, environmental impacts, and supply chain data).

Such reporting should be coupled with early or priority access to models for research and safety purposes, as recently offered by leading labs in the UK.[25] Mandatory disclosure and early access for safety would reduce information asymmetry and offer regulators prior warning of advancements in 'state of the art' capabilities and the ability to better prepare for the impact of these developments.[26]

Third-party conformity assessment

Another essential mechanism that the AI Act must introduce is third-party conformity assessment for foundation models. Self-assessments (first-party) and contracted auditing (second-party) have consistently been proven to be lower quality than accredited third-party or governmental audits.[27]

---

22 Kris Shrishak (n 12)

23 Sean McGregor, 'Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database' (arXiv, 17 November 2020) http://arxiv.org/abs/2011.08512 accessed 30 August 2023

24 Nikhil Mulani and Jess Whittlestone, 'Proposing a Foundation Model Information-Sharing Regime for the UK' (GovAI Blog, 16 June 2023) https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk accessed 30 August 2023

25 'PM London Tech Week Speech: 12 June 2023' (GOV.UK, 12 June 2023) https://www.gov.uk/government/speeches/pm-london-tech-week-speech-12-june-2023 accessed 16 June 2023

26 Ada Lovelace Institute (n 11)

27 Nopmanee Tepalagul and Ling Lin, 'Auditor Independence and Audit Quality: A Literature Review' (2015) 30 Journal of Accounting, Auditing & Finance 101; Bennett VM and others, 'Customer-Driven Misconduct: How Competition Corrupts Business Practices' (2013) 59 Management Science 1725

Given this evidence, third-party audits for foundation models have been one of the most widely suggested policy proposals across the AI ecosystem.[28 29 30 31 32]

Beyond safety and efficacy, third-party audits provide stronger guarantees and ease of compliance for the thousands of downstream users who build on top of or deploy the foundation model. For these reasons, we see third-party audits as one of the key risk-reducing mechanisms available to regulators, and so propose that the AI Act should compel foundation models to undergo third-party conformity assessment.

Pre-market third-party conformity should be assessed by notified bodies and supported by accredited third-party auditors. These audits should be based on full API and data access, and standardised assessment.[33] Throughout, the burden of proof should be on the foundation model developer to prove safety and efficacy, given that this is – for now – where the most expertise about the specific technology resides. Over time, this will also enable upskilling amongst auditors and regulators, similarly to the USA's Federal Drug Administration (FDA) learning and deepening its expertise via similar pre-market visibility and approval mechanisms.[34]

Structured access

In line with the AI lifecycle, following assessment of conformity, the question of responsible release and access should be addressed. The AI Act should compel staged release of foundation models based on 'structured access', whereby limits are placed on a system's use, modification and reproduction.[35] If serious issues or risks to people and society arise at this stage, full release on to the market should be postponed or blocked.

28 'One of the "Godfathers of AI" Airs His Concerns' The Economist https://www.economist.com/by-invitation/2023/07/21/one-of-the-godfathers-of-ai-airs-his-concerns accessed 30 August 2023

29 'Analyzing the European Union AI Act: What Works, What Needs Improvement' https://hai.stanford.edu/news/analyzing-european-union-ai-act-what-works-what-needs-improvement accessed 30 August 2023

30 'Zero Trust AI Governance' (Accountable Tech) https://accountabletech.org/research/zero-trust-ai-governance-framework/ accessed 30 August 2023.

31 Jakob Mökander, 'Auditing Large Language Models: A Three-Layered Approach' [2023] AI and Ethics http://arxiv.org/abs/2302.08500 accessed 30 August 2023

32 'AI Accountability Policy Request for Comment | National Telecommunications and Information Administration' https://www.ntia.gov/issues/artificial-intelligence/request-for-comments accessed 30 August 2023

33 'Zero Trust AI Governance' (n 30)

34 Ada Lovelace Institute (n 19).

35 Toby Shevlane, 'Structured Access: An Emerging Paradigm for Safe AI Deployment' (arXiv, 11 April 2022) <http://arxiv.org/abs/2201.05159 accessed 31 August 2023

This approach would be much safer than the current 'beta' testing on the public (as seen with the release of ChatGPT). This is proportionate, given that the most advanced foundation models are complex, novel and have difficult-to-foresee risk profiles – making it very difficult to know *ex ante* how they will operate in new environments.[36]

AI researchers have also suggested this could mitigate risks stemming from open-sourcing foundation model parameters and the inability to decommission misuse in such a release. An initial closed-source or structured-access release would allow a risk observation period, with the option to release model parameters once the regulator is confident in the risk-management framework.[37] Leading economists have advocated such an approach, citing the possibility for AI harms to be irreversible.[38] Beyond the risk-reducing features of such an approach, it has also been argued that rapid development and deployment would pose significant challenges for regulators in terms of enforceability of legislation.[39]

It is therefore better to get insights on how these models affect smaller numbers of people and slowly scale up than to test on the whole EU market (for example, ChatGPT reached 100 million users while still in beta testing)[40]. We recommend a more precautionary approach to rolling out foundation models across the economy.

A staged release based on structured access would also address the 'external access problem'.[41] This relates to concerns from AI developers around exposure to unnecessary privacy, security and intellectual property (IP) risk. Structured access would reduce these risks, while also facilitating an ecosystem of assessment (see 'Mitigating risk throughout the AI system lifecycle via an ecosystem of inspection').

Post-market monitoring

Once the foundation model is distributed, further risks can arise as the model can learn throughout its lifecycle, or be deployed in new areas. We therefore recommend post-market monitoring mechanisms.

36 Ada Lovelace Institute (n 19)

37 Anthony M. Barrett and others 'AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models' (forthcoming)

38 Daron Acemoglu and Todd Lensman, 'Regulating Transformative Technologies' (National Bureau of Economic Research, July 2023) https://www.nber.org/papers/w31461 accessed 30 August 2023

39 Shin-Shin Hua and Haydn Belfield, 'Effective Enforceability of EU Competition Law Under AI Development Scenarios: A Framework for Anticipatory Governance', Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (ACM 2023) https://dl.acm.org/doi/10.1145/3600211.3604694 accessed 1 September 2023

40 'ChatGPT and the rise of conversational AI models' https://www.peren.gouv.fr/rapports/2023-04-06_Eclairage_sur_CHATGPT_EN.pdf accessed 30 August 2023

41 'How to Audit an AI Model Owned by Someone Else (Part 1)' (OpenMined Blog, 30 June 2023) https://blog.openmined.org/ai-audit-part-1/ accessed 31 August 2023

First, there should be 'requirements to maintain an easy complaint mechanism for users to swiftly report any serious risks that have been identified'.[42] Once notified, the provider should take corrective action, document the incident and notify the relevant supervisory authority. This is particularly important for foundation models that are provided via an API, as in this case the provider maintains a significant degree of control over the underlying model,[43] meaning the provider will usually be in a position to mitigate or correct accidents or misuse.[44] It would also have the benefit of reducing burdens on regulators to open investigations or document incidents.

Second, foundation model providers and (subsequently) downstream application providers should label and make clear the underlying foundation model to end users and affected persons. This should include whether it is an open-source model, as well as easily accessible explanations of systems' main parameters and any opt-out mechanisms or available human alternatives.[45] This is important to ensure transparency for post-market tracking of harms and risks, and a prerequisite for a comprehensive remedies framework (see 'Redress and remedies').

Considering a 'tiered' approach for foundation models

Beyond the AI Act's obligations for foundation models, it is also necessary to consider their societal impact, which has the potential to be transformative – dependent on, for example, advancements in capabilities and the scale of their deployment. This might mean that measures for designating further obligations (on top of the European Parliament's Article 28b) for the most significant models could be considered, but it remains to be seen how this should be defined.

Suggested measurement metrics include expenditure on training runs; compute (measured in floating-point operations per second or 'FLOPS') used for both training and inference; capabilities evaluations; and post-deployment thresholds such as recipients of the model (for example, the number of API or enterprise customer providers); number of users of the service (for example, over 45 million or a number equivalent to 10% of the EU population); or the number of high-risk systems that build on the model.

The most future-proofed measure would be to empower the AI Office to issue binding guidance for particular strategically important foundation models, or for the AI Act to compel secondary legislation on this question. If thresholds are set to define strategically important foundation models, these criteria should be reviewed and updated every 12 months, and the AI Office should consult with all relevant stakeholders when doing so.

42 'Zero Trust AI Governance' (n 30)

43 Sabrina Küspert , Nicolas Moës and Connor Dunlop (n 15)

44 'Governing General Purpose AI — A Comprehensive Map of Unreliability, Misuse and Systemic Risks' (20 July 2023) https://www.stiftung-nv.de/de/publikation/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks accessed 30 August 2023

45 'Zero Trust AI Governance' (n 30)

In the same way that additional requirements may need to be considered for the most powerful or most widely deployed models, there might also be a case for reduced obligations on foundation models that are deemed to have some public benefits. This may include some open-source foundation models (open-source developers have suggested they should only comply with some elements of the European Parliament's proposed Article 28b)[46] but it will be important to distinguish between 'true' open-source models and those that are provided for commercial gain and/or ecosystem capture.[47]

Open-source foundation models

The question of 'open-source' foundation models is a key one for the trilogues. Historically, open-source software was usually a non-commercial endeavour, often in the pursuit of science. However, the relationship of 'open source' to the technology industry has fostered complex incentives and often contradictory rhetoric of openness. Examples include Google's release of Android as an open-source operating system as a means to control a platform, and more recently Meta's provision of PyTorch as an open-source framework to enable the company to integrate open-source products into its proprietary systems.[48]

There are some signs of similar incentives and contradictory rhetoric in the provision of foundation models. Given its historical connotations, open-source AI is often thought of as similar to open-source software. However, it is important to recognise that this is not accurate, as the 'resources needed to build AI from scratch, and to deploy large AI systems at scale, remain "closed" – available only to those with significant (almost always corporate) resources'.[49] In addition, even when foundation models are provided via 'open source', they will often be behind an interface (API) and/or indirectly monetised (for example, through selling associated services or 'closing' an advanced version of the model).[50]

We recommend that it would therefore be safest to take a precautionary approach, adopting the European Parliament's suggestion to regulate foundation models regardless of distribution channel. The AI Office should be empowered to designate exemptions for open-source models that are shown to be built for public benefit (rather than commercial, for example, collaborative development of free open source[51]), such as models built for governments or open-source collectives. This would be a more future-proofed approach than blanket exemptions, particularly

46 Peter Cihon, 'How to Get AI Regulation Right for Open Source' (The GitHub Blog, 26 July 2023) https://github.blog/2023-07-26-how-to-get-ai-regulation-right-for-open-source/ accessed 30 August 2023

47 David Gray Widder, Sarah West and Meredith Whittaker, 'Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI' (17 August 2023) https://papers.ssrn.com/abstract=4543807 accessed 30 August 2023

48 Meta CEO Mark Zuckerberg has described how open sourcing the PyTorch framework has made it easier to capitalize on new ideas developed externally and for free. See also: David Gray Widder, Sarah West and Meredith Whittaker (n 47)

49 David Gray Widder, Sarah West and Meredith Whittaker (n 47)

50 Sabrina Küspert , Nicolas Moës and Connor Dunlop (n 15)

51 'European Parliament Gives Green Light to AI Act, Moving EU towards Finalizing the World's Leading Regulation of AI' (Creative Commons, 14 June 2023) https://creativecommons.org/2023/06/14/european-parliament-gives-green-light-to-ai-act-moving-eu-towards-finalizing-the-worlds-leading-regulation-of-ai/ accessed 3 September 2023.

as some have suggested that open-source exemptions could increase regulatory burdens for downstream SMEs and startups who build on top of open-source models.[52]

Regulatory alignment

Our proposal to include independent audits, user feedback and complaint mechanisms would also be helpful for the EU in terms of regulatory alignment with other jurisdictions. In the USA, the White House has announced a commitment (voluntary for now, Executive Order pending) for (frontier) foundation model developers to facilitate 'third-party discovery and reporting of vulnerabilities in their AI systems', as well as a 'robust reporting mechanism' to allow 'issues (that) may persist even after an AI system is released' to be 'found and fixed quickly'.[53] These interventions hold significant overlap with independent audits and user reporting of misuse or other issues, and adopting them would mean the EU's regime would align better with the USA's approach to foundation model governance.

Legal loopholes

Finally, the (potential) legal loophole for general-purpose AI in the Council of the EU's general approach (Article 4c) should be closed in the final AI Act text. Article 4c(1) appears to allow developers to relinquish responsibility using a standard legal disclaimer. Such an approach creates a dangerous loophole that reduces the responsibilities on original developers of foundation models/GPAI (who are often well-resourced companies), and instead places sole responsibility with downstream actors that lack the resources, access and ability to mitigate all risks.

52 Some have claimed open source exemptions could in fact increase, rather than reduce, the burden on small businesses. Exempting open-source models from regulatory scrutiny could create an undue competitive burden for smaller firms, by ensuring that compliance requirements fall on the entity that fine-tunes an end product.

53 The White House, 'FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI' (The White House, 21 July 2023) https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/ accessed 1 September 2023

Recommendations:

- The European Parliament's proposed rules for foundation model (Article 28b) should be implemented in full and strengthened further (detailed in the following bullet points). This would represent a comprehensive regime that would allay the concerns of European SMEs and civil society around 'smaller providers and users bearing the brunt of obligations better suited to original developers'.[54] [55]

- Mandatory disclosure requirements for developers of foundation models operating in the EU should be introduced, including notification to the AI Office or national regulators when beginning large-scale training runs, coupled with disclosure of compute and capabilities evaluations.[56] This disclosure should also include the high-value information proposed by the European Parliament in Annex VIII, Section C. Regulators and vetted researchers should also be granted early access to foundation models to increase visibility over risks.

- The AI Act should introduce pre-market, third-party conformity assessment for foundation models, ensuring the burden of proof is on developers to prove safety and compliance before widescale distribution. Conformity assessments should be based on assessment of the quality management system and technical documentation, with the involvement of a notified body, referred to in Annex VII – not on internal controls. Regulators should use independent experts or auditors to facilitate this process if there is an initial lack of regulatory expertise for such external scrutiny.

- Foundation models should be rolled out in a staged manner, and provided via 'structured access' – at least until regulators have more visibility over how these models interact with people and in complex environments.

- Developers of foundation models provided via an API should offer a complaint mechanism for users to swiftly report any serious risks that have been identified. The provider should take corrective action when possible and notify the relevant supervisory authority when not possible.

54 Giorgos Verdi, 'General-Purpose AI Fit for European Small-Scale Innovators' (European DIGITAL SME Alliance, 5 October 2022) https://www.digitalsme.eu/general-purpose-ai-fit-for-european-small-scale-innovators/ accessed 1 September 2023

55 'EU Trilogues: The AI Act must protect people's rights A civil society statement on fundamental rights in the EU Artificial Intelligence Act' https://edri.org/wp-content/uploads/2023/07/Civil-society-AI-Act-trilogues-statement.pdf accessed 30 August 2023

56 Nikhil Mulani and Jess Whittlestone (n 24)

- Foundation model providers and downstream application providers should label and make clear the underlying foundation model to end users and affected persons. This should include whether it is an open-source model, as well as easily accessible explanations of systems' main parameters and any opt-out mechanisms or available human alternatives.

- Foundation models should be regulated regardless of distribution channel, meaning open-source exemptions should not be applied to this specific type of AI. The AI Office should be empowered to add additional obligations for foundation models deemed uniquely capable or strategically significant, or to issue partial exemptions for models deemed to be of public benefit (for example, collaborative development of free open source). If thresholds are set to define a sub-category of foundation models, this criteria should be reviewed and updated every 12 months, and the AI Office should consult with all relevant stakeholders when doing so.

- The legal loophole for general-purpose AI in the Council of the EU's general approach (Article 4c) should not be included in the final AI Act, so that developers cannot relinquish responsibility using a standard legal disclaimer.

3. Mitigating risk throughout the AI system lifecycle via an ecosystem of inspection

Pre-market and pre-deployment conformity is an essential component to ensuring the AI Act works to protect people and society. However, inspection cannot stop at this point. Other sectors have shown that 'one and done' conformity checks have the potential to be gamed or to miss emergent behaviours (see the Volkswagen emissions scandal).[57] These issues are even more likely for AI, given its dynamic nature, including the capacity to change throughout the lifecycle and for downstream users to fine-tune and (re)deploy models.

Vetted researcher access

Continuous monitoring and auditing of systems placed on the market will therefore be needed. This will require fostering of an ecosystem of assessment. A key element of this ecosystem should be access to foundation models for researchers, and the ability to carry out external inspections (for example, 'red-teaming', second- and third-party audits).[58] [59]

57 Bill Chappell, "It Was Installed For This Purpose," VW's U.S. CEO Tells Congress About Defeat Device' NPR (8 October 2015) https://www.npr.org/sections/thetwo-way/2015/10/08/446861855/volkswagen-u-s-ceo-faces-questions-on-capitol-hill accessed 30 August 2023

58 Ada Lovelace Institute (n 16)

59 Ada Lovelace Institute, Technical methods for regulatory inspection of algorithmic systems (2023) https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/ accessed 30 August 2023

This is beneficial as external researchers have different incentives (and expertise) in relation to finding vulnerabilities than developers, regulators or auditors. Our research has found that such an ecosystem can help overcome the challenges of finding adequate capacity for technical oversight.[60] while the Stanford Institute for Human-Centered Artificial Intelligence argues that no organisation can be equipped to anticipate all risks because foundation models can be adapted to multiple downstream applications, and therefore 'it is imperative that external researchers representing a diversity of institutions, cultures, demographic groups, languages, and disciplines be able to critically examine foundation models from different perspectives'.[61]

To make such an ecosystem viable, foundation models should be made available via structured access (see section on 'Structured access') and vetted researchers should also be granted 'fair use protections' in cases where they may violate Terms of Service in external audit investigations.[62]

This was suggested by a high-level expert group who called for 'qualified researchers and auditors who meet certain conditions (to) be given model-and-system framework access', coupled with 'the establishment of narrowly-scoped "safe harbour" provisions for industry and researchers, designed to reasonably assure that entities participating in good faith auditing exercises are not subjected to undue liability risk or retaliation.' [63]

Benchmarking

To ensure effective scrutiny and uniformity of application, benchmarking and measurement will be essential. The European Commission's Joint Research Centre's AI Watch has found that private performance benchmarks, competitions and challenges are behind much of the recent progress in AI.[64]

Without effective measurement and visibility over the most advanced models, it is very hard to understand if and when development and deployment is safe. We therefore support the European Parliament's proposal for the AI Office to work with international metrology (the science of measurement in research and innovation) and benchmarking authorities to develop metrics for safe AI development and deployment (as proposed in Article 58a).

60 Ibid.

61 'The Time Is Now to Develop Community Norms for the Release of Foundation Models' (Stanford HAI) https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models accessed 30 August 2023

62 See: Ada Lovelace Institute (n 59): 'In some jurisdictions there have been legal challenges to scraping audits and concerns as to whether the act of web scraping violates platforms' terms and conditions [...] Regulators would avoid this concern by having explicit powers to conduct this work, and could help in clarifying the rights of others to perform scraping audits and encouraging openness of platforms to these approaches.'

63 'AI Accountability Policy Request for Comment | National Telecommunications and Information Administration' https://www.ntia.gov/issues/artificial-intelligence/request-for-comments accessed 30 August 2023

64 Fernando Martınez-Plumed, Jose Hernandez-Orallo and Emilia Gomez, 'Tracking the Impact and Evolution of AI: The AIcollaboratory'

However, we would propose that this should go further by setting up a Benchmarking Institute, as first proposed in the European Parliament's ITRE Committee.[65] Our research and a subsequent expert roundtable identified this as one of the most effective ways to support standards-setting processes and the wider auditing ecosystem.[66] [67] A Benchmarking Institute could be funded by a mandatory fee paid by developers over a certain threshold (to be defined by the AI Office), since such measurement can also contribute to improving performance, on top of enhancing safety.[68]

The development of an approach and ecosystem of independent experts who can work on AI audits and external inspection will take time. This challenge has been met in other sectors, from cybersecurity and civil aviation to automotives and financial services.[69]

For example, the civil aviation sector used an ecosystem approach to monitor and document incidents, and proactively offer risk-mitigation strategies. This helped encourage a culture of safety in the industry, which reduced fatality risk by 83% between 1998 and 2008 (while seeing a 5% annual increase in passenger kilometres flown).[70] In addition, it should also be mentioned that many organisations already exist in this space to offer auditing services (for example, Eticas AI, AppliedAI, Algorithmic Audit, and Apollo research), and many more will continue to be set up. [71]

65 Draft Opinion of the Committee on Industry, Research and Energy, European Parliament, 3.3.2022

66 Ada Lovelace Institute, Inclusive AI governance (2023) https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/ accessed 30 August 2023

67 'EU AI Standards Development and Civil Society Participation' (Ada Lovelace Institute) https://www.adalovelaceinstitute.org/event/eu-ai-standards-civil-society-participation/ accessed 30 August 2023

68 'Written Testimony of Jack Clark Co-founder, Anthropic. Co-chair, AI Index. Member, National AI Advisory Committee. Before the U.S. Senate Committee on Commerce, Science, and Transportation Thursday September 29th, 2022' https://www.commerce.senate.gov/services/files/F7BFA181-1B1B-4933-A815-70043413A7FF accessed 30 August 2023

69 Inioluwa Deborah Raji and others, 'Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance', Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (ACM 2022) https://dl.acm.org/doi/10.1145/3514094.3534181 accessed 30 August 2023

70 Abhishek Gupta, 'Emerging AI Governance Is an Opportunity for Business Leaders to Accelerate Innovation and Profitability' (Tech Policy Press, 31 May 2023) https://techpolicy.press/emerging-ai-governance-is-an-opportunity-for-business-leaders-to-accelerate-innovation-and-profitability/ accessed 30 August 2023

71 'Key Enforcement Issues of the AI Act Should Lead EU Trilogue Debate' (Brookings) https://www.brookings.edu/articles/key-enforcement-issues-of-the-ai-act-should-lead-eu-trilogue-debate/ accessed 30 August 2023

Recommendations

- Developers and deployers should offer vetted researcher access to carry out external inspections – such as red-teaming – to test AI models for vulnerabilities and risks to health, safety or fundamental rights – as included in the EU Digital Services Act.
- The AI Act should include 'safe harbour' provisions for industry and researchers, designed to reasonably assure that entities participating in good-faith auditing exercises on a good-faith basis are not subjected to undue liability risk or retaliation.
- An EU Benchmarking Institute should be set up to coordinate benchmarking and national metrology authorities – and other relevant stakeholders – on creating benchmarks and thresholds for measuring safe development and deployment of AI models. This could be partly funded by mandatory contributions from AI developers over a certain threshold (to be decided by the AI Office). This will support both standards bodies and the EU's auditing and assurance ecosystem, help with uniformity in safety and compliance, and ultimately trust and uptake in AI in the EU single market.

4. Maintaining a risk-based approach and ensuring future-proofed regulation

'Filters' on high-risk categorisation

We share civil society concerns regarding the European Parliament and the Council of the EU's suggestions to narrow the scope of the Act by introducing an additional 'filter', which means high-risk systems are no longer determined by the area in which they are deployed.[72]

It is not clear how determination of either 'substantial risk' or 'purely accessory' decisions would be done, and it may undermine the effectiveness of the whole regulation to allow this determination to be made by AI developers and deployers themselves.

If there is to be a 'filter', it should be based on a precautionary approach, as in the European Commission's proposal for classifying high-risk. This should be coupled with empowering the AI Office or the European Commission to issue guidelines to remove certain use cases when there is evidence that they do not pose risks to people and society.

72 'EU Trilogues: The AI Act must protect people's rights A civil society statement on fundamental rights in the EU Artificial Intelligence Act' https://edri.org/wp-content/uploads/2023/07/Civil-society-AI-Act-trilogues-statement.pdf accessed 30 August 2023

Updating the list of 'high-risk' categories

In addition, the proposed Act currently assumes that the list of categories of high-risk AI systems in Annex III is comprehensive and complete. It acknowledges the need to add new uses of technologies, but only allows the European Commission to add new subcategories (for example, subcategories of foundation models). This brings partial futureproofing that is out of step with the nature of AI, which evolves quickly and significantly.

A mechanism should therefore be put in place to ensure that the list of categories in Annex III can be extended, as well as retaining the ability to add sub-categories below those existing categories (for high-risk uses and also sub-categories of foundation models).[73] This mechanism could be informed by incident reporting and the AI Office's monitoring and foresight activities, as well as individual complaints brought by affected persons (see 'Redress and remedies').

The European Commission should also enable the public to express their concerns and flag any systems that they believe should be added to this list. The European Commission should have an obligation to consider these concerns (and where relevant, together with the AI Office) in a timely manner and present a reasoned response.

Criteria for risk categorisation

It is unclear what criteria has been used in placing different AI systems into different risk categories. Some examples of AI appear to have been miscategorised, for example deep fakes and emotion recognition are placed in the limited-risk category, despite the systemic risks they pose – from misinformation to gender hate.[74]

Criteria for risk categorisation would clarify the rationale behind categorisation, and should itself be open to scrutiny and challenge, for example via a permanent sub-group of the AI Office Advisory Forum. This would strengthen confidence in categorisation of current AI systems, and further future-proof the Act to enable appropriate categorisation of new uses.

Beyond clarifying these criteria, they should also be expanded beyond just health, safety and fundamental rights to include systemic and environmental risks.[75] Systemic harms from AI may arise and have the potential to alter the dynamics of social, political, economic and environmental systems (such as the potential for jobs to change or be significantly altered as a result of AI automation or augmentation, or the aggregate effect of misinformation on democratic institutions).[76]

---

73 Ada Lovelace Institute, 'People, Risk and the Unique Requirements of AI' https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act/ accessed 30 August 2023

74 Ibid

75 Ada Lovelace Institute (n 66)

76 Ada Lovelace Institute (n 11)

These harms are often overlooked in the context of AI, with measures designed to address individual and collective harm – but these measures are not always suitable to counter systemic risks. We would therefore encourage an approach to AI that addresses its effects on society.[77] [78] For the AI Act, this would merit the inclusion of language on systemic and environmental risks, empowering regulators to update risk profiles based on these if and when deemed necessary.

Recommendations

- New filters should not be introduced for high-risk categorisation. The European Commission's proposal for high-risk categorisation should be maintained, possibly coupled with a case-precedent regime in which the AI Office is empowered to issue guidance to remove certain use cases from high-risk designation.
- A clear mechanism for updating the high-risk list of categories in Annex III should be included, as well as retaining the ability to add sub-categories below those existing categories. Affected persons should be empowered to flag any systems that they believe should be added to this list.
- Any changes to the high-risk list should be based on clear, judicially reviewable criteria, and should consider systemic and environmental risk.

5. Protection and representation for affected persons

Definitions

To accurately reflect the AI lifecycle and include those who are ultimately affected by the deployment of an AI system within the framework laid down by the Act, we propose including definitions of 'affected persons' and renaming 'users' to 'deployers'.[79]

77 Nathalie A. Smuha, 'Beyond the individual: Governing AI's societal harm' 10(3) Internet Policy Review https://policyreview.info/articles/analysis/beyond-individual-governing-ais-societal-harm accessed 30 August 2023

78 'Governing General Purpose AI — A Comprehensive Map of Unreliability, Misuse and Systemic Risks' (20 July 2023) https://www.stiftung-nv.de/de/publikation/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks accessed 30 August 2023

79 Ada Lovelace Institute (n 73)

This will better distinguish those who deploy systems created by providers and those who use or are ultimately affected by the use of AI systems. For example, students whose final grades are determined by an AI-based system would be affected persons, as would job applicants whose CVs are processed by an AI-based system.[80] We therefore support the European Parliament's inclusion of a definition of affected persons and suggestion to use the language of 'deployer'.

Pre-deployment impact assessments

To offer protections for affected persons, it will also be necessary to include algorithmic impact assessments,[81] or fundamental rights impact assessments (FRIA) as proposed by the European Parliament. This is because AI development and deployment will have societal and human rights implications (for example, in relation to the 'accuracy' of a system – what is an acceptable level of false positives for an AI system recommending criminal sentences in the justice system?).

Fundamental rights will therefore have to be considered when deploying an AI system, and a fundamental rights impact assessment as proposed by the European Parliament represents the best way to ensure these implications are considered.

Redress and remedies

A FRIA is not a catch-all solution for (post-)deployment however. There will inevitably be times when harm is caused, and people need strong redress and remedial protection to ensure safety and trust in the AI ecosystem. We therefore echo leading AI researchers in strongly welcoming the avenues for remedy and redress introduced in Chapter 3a of the European Parliament's text, such as a right to lodge a complaint with a supervisory authority, judicial remedy and an explanation of individual decision-making.[82] These must be included in the final Act, to offer recourse for (harmed) people and also to complement the forthcoming AI Liability Directive.

Multi-stakeholder participation

When considering fundamental rights protections, it is also worth considering the AI Act's reliance on technical standards to provide the detailed guidance necessary for compliance. Standards development bodies lack the expertise and legitimacy to make decisions with fundamental rights implications, of which there will be many.[83] This misalignment is significant because it has the potential to leave fundamental rights and other public interests unprotected.

80 Ada Lovelace Institute (n 66)

81 'Algorithmic Impact Assessment: User Guide' (Ada Lovelace Institute, 2023) https://www.adalovelaceinstitute.org/resource/aia-user-guide/ accessed 30 August 2023

82 Meeri Haataja and Joanna J Bryson, 'The European Parliament's AI Regulation: Should We Call It Progress?' (2023) 4 Amicus Curiae 707.

83 Hadrien Pouget, 'What Will the Role of Standards Be in AI Governance?' (Ada Lovelace Institute, 5 April 2023) https://www.adalovelaceinstitute.org/blog/role-of-standards-in-ai-governance/ accessed 30 August 2023

We therefore support the European Parliament's suggestions for enhancing multi-stakeholder participation through new provisions on the participation of stakeholder groups in standards development processes and on enhancing transparency over the processes.[84] The aim would be to achieve 'a balanced representation of interests by involving all relevant stakeholders in the development of standards' (Recital 61).

We also support the amendment to Article 40 that specifies that, while drafting the standardisation request, the European Commission will consult with the AI Office and its Office's Advisory Forum (a body responsible for providing the AI Office with inputs from different stakeholder groups).

This provision could go further still, however. A recent academic paper concluded that the main way the AI Act can be strengthened is with more 'effective citizen engagement'.[85] One mechanism for this would be to have a standing panel of representative users – a type of 'citizens assembly' as suggested in the aforementioned report – as a permanent sub-group of the Office's Advisory Forum. Ada's Citizens' Biometrics Council[86] (in the UK) has been highlighted as a leading example of this type of engagement, and we would like to work with policymakers to explore if and how such an approach can be developed under the AI Act.

This assembly could be a mechanism to elicit feedback from affected persons in AI questions with societal-level implications (for example, release of large-scale models and areas for their deployment), which has been suggested by AI labs. OpenAI and Anthropic have participated in 'alignment assemblies', which seek to use public opinion to inform criteria for responsible release of models, for example. While experiments of this sort are valuable, we would recommend formalising this through public participation mechanisms that also have regulatory oversight, and that permanent sub-groups of the AI Office could be an effective means to do so.[87]

84 Arcangelo Leone de Castris and Chris Thomas, 'What Role Do Standards Play in the EU AI Act? Looking at the Implications of the European Parliament's Proposed Amendments' (AI Standards Hub, 24 July 2023) https://aistandardshub.org/eu-ai-act/ accessed 30 August 2023

85 Huw Roberts and others, 'Governing Artificial Intelligence in China and the European Union: Comparing Aims and Promoting Ethical Outcomes' (2023) 39 The Information Society 79.

86 Ada Lovelace Institute, The Citizens' Biometrics Council (2021) https://www.adalovelaceinstitute.org/report/citizens-biometrics-council/ accessed 5 September 2023.

87 'Alignment Assemblies' (The Collective Intelligence Project) https://cip.org/alignmentassemblies accessed 30 August 2023/

Recommendations

- The AI Act should include a definition of 'affected persons' and rename 'users' to 'deployers' to more accurately reflect the AI lifecycle.

- Include an obligation for a pre-deployment fundamental rights impact assessment (FRIA), as proposed by the European Parliament.

- Include a comprehensive remedies framework for affected persons based on effective avenues for redress, including a right to lodge a complaint with a supervisory authority, judicial remedy and an explanation of individual decision-making – as proposed in Chapter 3a of the European Parliament's text.

- Enhance multi-stakeholder participation in standards development processes and enhance transparency over these processes. The European Commission should consult with the AI Office's Advisory Forum when drafting standardisation requests, or approving harmonised standards.

- Include a standing panel of representative users or a citizens' assembly as a permanent sub-group of the Office's Advisory Forum. The panel should be consulted on key questions such as updating the high-risk list, release of large-scale foundation models, or secondary legislation.

Ada's work related to the EU AI Act

- *Regulating AI in Europe: four problems and four solutions*: An expert opinion paper by Professor Lilian Edwards https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/
- 'People, risk and the unique requirements of AI': Our 18 recommendations for strengthening the AI Act in early 2022. https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act/
- Explainer: 'What is a foundation model?' https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/
- 'The value chain of general-purpose AI': A closer look at the implications of API and open-source accessible GPAI for the EU AI Act https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/
- *Allocating accountability in AI supply chains:* An expert explainer by Professor Ian Brown https://www.adalovelaceinstitute.org/resource/ai-supply-chains/
- *AI assurance? Assessing and mitigating risks across the AI lifecycle* https://www.adalovelaceinstitute.org/report/risks-ai-systems/
- *Inclusive AI governance: Civil society participation in standards development* https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/
- *Keeping an eye on AI: Approaches to government monitoring of the AI landscape* https://www.adalovelaceinstitute.org/report/keeping-an-eye-on-ai/

Ada Lovelace Institute
100 St John Street, London, WC1B 3JS
+44 (0) 20 7631 0566

Registered charity 206601

Website: adalovelaceinstitute.org
Twitter: @AdaLovelaceInst
Email: hello@adalovelaceinstitute.org