# Five considerations to guide the regulation of "General Purpose AI" in the EU's AI Act:

## Policy guidance from a group of international AI experts

The hype surrounding the high-profile release of multiple Large Language Models[1], Image generators, and other generative AI technology like DALL-E or Midjourney in the last few weeks and months has brought renewed urgency to the question of whether (and how) so-called "general purpose AI" (GPAI) should be regulated. In Europe, this question is not hypothetical. Rather it is the subject of a hotly contested[2] policy debate around the Artificial Intelligence Act, the EU's flagship AI regulation, which has been evolving for close to two years now.

Introduced by the European Commission in April 2021, the Commission's original AI Act proposal effectively exempted the developers of GPAI from complying with a range of documentation and other accountability requirements in the law.[3] These requirements only apply to high-risk AI which is defined in the Act based on use/context. This would therefore mean that GPAI that ostensibly had no predetermined use or context would not qualify as 'high risk' – another provision (Article 28) confirmed this position, implying that developers of GPAI would only become responsible for compliance if they significantly modified or adapted the AI system for high-risk use. The European Council's subsequent "general approach" for trilogue[4] negotiations took a different stance where original providers of GPAI will be subject to certain requirements in the law, although working out the specifics of what these are or should be would be delegated to the Commission. Recent reports suggest that the European Parliament, too, is considering obligations specific to original GPAI providers.

**The AI Act's approach to general purpose AI is poised to set the regulatory tone for addressing AI harms globally** (not unlike the regulatory momentum set in motion by Europe's data protection law, the GDPR). With the recent upswell in public attention on generative AI, there's also a risk that the regulatory position ends up overfitting to the concerns of the day. **But contemporary technologies such as ChatGPT, DALL-E 2, and Bard are just the tip of the iceberg.** In this brief, we recommend clear definitional signposts and principles to ensure that the approach in the AI Act is as future proof as possible. As a group of international AI experts across domains of computer science, law

[1] Like ChatGPT and its subsequent adaptation into Microsoft's search chatbot; Anthropic and Google's announcements around their respective models 'Claude' and 'Bard'

[2] Corporate Europe Observatory, "The Lobbying Ghost in the Machine: Big Tech's covert defanging of Europe's AI Act", February 2023, https://corporateeurope.org/sites/default/files/2023-02/The%20Lobbying%20Ghost%20in%20the%20Machine.pdf

[3] 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," April 21, 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206.

[4] In the EU legislative process, trilogue negotiations between EU institutions — that is, the Council of the EU, the European Parliament, and the European Commission — are the final step before a legislative proposal can be adopted.

and policy, and the social sciences, we propose the following considerations to guide the EU's regulatory approach to AI. While there are likely deeper concerns with the law's substantive approach (for e.g. the delegation to standard-setting organizations has been heavily criticized[5]), this comment restricts itself to questions like how should GPAI be understood and categorized? What are the inherent risks associated with GPAI models? What is the relationship between these and downstream harms that emerge based on specific use cases? Whether and at what stage/s do GPAI models require regulatory guardrails, oversight and scrutiny? Based on these enquiries, we propose five key considerations to guide the regulation of GPAI:

1. **GPAI is an expansive category.** For the EU AI Act to be future proof, it must apply across a spectrum of technologies, rather than be narrowly scoped to chatbots/large language models (LLMs). The definition used in the Council of the EU's general approach for trilogue negotiations provides a good model.
2. **GPAI models carry inherent risks and have caused demonstrated and wide-ranging harms.** While these risks can be carried over to a wide range of downstream actors and applications, they cannot be effectively mitigated at the application layer.
3. **GPAI must be regulated throughout the product cycle, not just at the application layer, in order to account for the range of stakeholders involved.** The original development stage is crucial, and the companies developing these models must be accountable for the data they use and design choices they make. Without regulation at the development layer, the current structure of the AI supply chain effectively enables actors developing these models to profit from a distant downstream application while evading any corresponding responsibility.
4. **Developers of GPAI should not be able to relinquish responsibility using a standard legal disclaimer.** Such an approach creates a dangerous loophole that lets original developers of GPAI (often well-resourced large companies) off the hook, instead placing sole responsibility with downstream actors that lack the resources, access, and ability to mitigate all risks.
5. **Regulation should avoid endorsing narrow methods of evaluation and scrutiny for GPAI that could result in a superficial checkbox exercise. This is an active and hotly contested area of research and should be subject to wide consultation, including with civil society, researchers and other non-industry participants.** Standardized documentation practice and other approaches to evaluate GPAI models, specifically generative AI models, across many kinds of harm are an active area of research. Regulation should avoid endorsing narrow methods of evaluation and scrutiny to prevent this from resulting in a superficial checkbox exercise.

[5] Michael Veale and Frederike Borgesius, Demystifying the Draft EU Artificial Intelligence Act, 4 Computer Law Review International (2021)

## 1. GPAI is an expansive category. For the EU AI Act to be future proof, it must apply across a spectrum of products, rather than narrowly scoped to chatbots/LLMs.

The definition used in the Council of the EU's general approach for trilogue negotiations is expansive in its definition of GPAI, and is currently defined as follows:

> *intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems.*

This definition is a step in the right direction, as it defines GPAI broadly and avoids several common pitfalls in this debate. **First, a suitable definition of GPAI ought to include many methods ("tasks") upon which other AI systems can be built:** This should include technologies of computer vision (e.g. facial analysis and recognition, object localization and recognition, scene segmentation), textual processing (e.g. machine translation, document summarization), automated speech recognition, and multi-modal analysis (generation of images from text prompts, reverse image search).

**Second, there is a further risk to narrowly bound the definition GPAI to novel image and text generative AI systems, such as OpenAI's DALL-E and GPT models.** Scoping the definition of GPAI to these systems would ignore a large class of models which could potentially cause significant harm if left unchecked, including many of the models already commonly offered via large cloud services such as Amazon Web Services (AWS) or Microsoft Azure. For example, it should also include AI systems which may be more mundane but can have deleterious downstream effects, such as complicated statistical models which can be tooled for uses such as predicting health risks from hospital data, or to allocate welfare benefits from administrative data. For the definition of GPAI to be as technologically neutral as possible, the focus should lie not on specific technologies underlying the latest innovations in this space, but rather on any AI model versatile enough to be adapted to a variety of different ("high-risk") uses within the meaning of the AI Act.

2. GPAI models carry inherent risks and have caused demonstrated and wide-ranging harms. The fact that these risks can be carried over to a range of downstream uses heightens, rather than mitigates, the need for regulation at the stage of development.

In considering regulatory approaches for GPAI, we ought to consider the current state of AI technologies, their uses and how they function. Currently, a subset of general purpose AI models can serve as the basis for many other currently conceptualized and imagined tasks, for which they are "fine-tuned." This includes patently "high risk" tasks such as AI for resume screening or assessing creditworthiness, but also comparably benign tasks such as using AI for generating poetry. Although the risks of downstream harms may appear to be at different scales for such distinct tasks, **there are several examples where risks are inherent in the use of GPAI models that serve as a basis for downstream tasks.**

For instance, GPAI models have been shown to carry the risk of producing anti-democratic speech such as hate speech targeted against gender, sexual, racial, and religious minorities. These models risk entrenching narrow or biased worldviews embedded in the underlying data. Furthermore, GPAI models have also shown that they memorize and reproduce private and personal information such as phone numbers, addresses, and medical documents. The potential harms are wide-ranging, including effects on privacy, representation, and access to social services that are dependent on protected traits such as race and gender. These harms will ensue regardless of how downstream actors retool the GPAI model, and they can only be addressed in the GPAI model itself.

These potential risks are inherently systemic in that they are likely to have an impact both at the individual and aggregate population level—and both are interconnected. There is a need to pay attention to the population-level effects of GPAI models as well as the individual ones. For example, the creation of a large-scale image dataset may be developed through mass surveillance, mass scraping of people's images without their consent, and cooperation with law enforcement agencies. These development practices contribute to the downstream issues where a facial recognition tool may mis-identify criminal suspects belonging to a particular racial minority. Although the individual-level harm is clear, the risks associated with the development of an GPAI model and its unaccounted systemic issues also need attention. How the model is built is an example of one such systemic concern. Moreover, any regulatory response or transparency mandate will need to involve multiple institutions and stakeholders.

Finally, **the fact that such GPAI models can be fine-tuned for specific uses and tasks only heightens the risk that the results would be unfair, inaccurate, or harmful in unanticipated ways.** The wide applicability of GPAI models to a range of tasks initially seems to reduce the risks of the GPAI models that undergird task-specific technologies, yet they carry many of the same risks to private individuals regardless of their particular use. **In fact, the inherent malleability of GPAI models is precisely why GPAI models, in addition to models for specific use-cases, must be regulated.** It may be difficult for any particular body or entity to imagine all use-cases and thereby correctly assess the level of risk for each use case, however, by considering the risks of technologies that are malleable to different tasks, and identifying vectors of risk, developers and regulatory bodies can address the risks to privacy and democratic participation prior to the advent of concerns that might arise. GPAI models represent a system of practices, stakeholders, impacts and relationships, and need a dynamic approach that takes all of these into account.

3. GPAI must be regulated throughout the product cycle, not just at the application layer, in order to account for the range of stakeholders involved. The original development stage is crucial, and the companies developing these models must be accountable for the data and design choices they make.

**In order for regulation to be effective, general purpose AI systems must be regulated throughout the entire product cycle and not solely at the application layer when they have been implemented for a particular purpose and context.** This includes data collection, cleaning, and annotation, as well as model development, testing, and evaluation. Excluding other points within the life cycle of GPAI leaves a massive gap in regulatory oversight which firms will be strongly incentivized to take advantage of:

- The implementation and eventual tooling of an AI system is an important point for regulatory intervention, but a sole focus on how GPAI systems are used at the application layer overlooks key aspects of AI development. For example, in most instances downstream actors (both users and downstream providers) lack the capacity to access and interpret core components of a pre-trained model – such as training data or evaluation gaps in the base model – while original developers can do so.

- Without regulation at the development layer, the current structure of the AI supply chain effectively enables actors developing these models to profit from a distant downstream application while evading any corresponding responsibility. This is even more concerning given that many GPAI developers are large companies with ample resources for conducting such evaluations, whereas downstream actors are more dispersed.
- After implementation, many GPAI systems retain structural dependencies which are tied to the originating developer, and excluding the pre-implementation stage wholly ignores risks related to these structural dependencies (for example, integration with the developer's cloud infrastructure, etc).
- Given this, the companies developing these models must be accountable for the data, modeling, and other design choices they make. Ignoring them will strongly incentivize AI developers to focus on releasing 'white-labeled' GPAI systems that evade regulatory scrutiny and offer a major loophole for liability.

4. Any regulatory approach that allows developers of GPAI to relinquish responsibility using a standard legal disclaimer would be misguided. It creates a dangerous loophole that lets original developers of GPAI (often well resourced large companies) off the hook, instead placing sole responsibility with downstream actors that lack the resources, access, and ability to mitigate all risks.

The Council's general approach, while including dedicated obligations for original developers of GPAI, comes with an exception that allows GPAI developers to rid themselves of responsibility as long as they "explicitly excluded all high-risk uses in the instructions of use or information accompanying the general purpose AI system" and and have no "sufficient reasons to consider that the system may be misused". Such an approach is misguided for several reasons:

- A legal disclaimer that is buried in technical documentation or instructions of use provides an efficient escape clause for large developers and disincentivizes due diligence of any kind at the

original developer stage. The failures of the “notice and consent” model for privacy are testament to the risks of using boilerplate language to enforce accountability. Privacy notices are usually ineffective in informing users of the ways in which their data is being used. This kind of informational asymmetry will likely be more pronounced with GPAI models that can be used in a lot more complex and multifaceted ways than user data that is collected by companies.

- Pre-trained GPAI models carry several inherent risks that stem from data and design choices made at this preliminary stage (see Point 2 and 3). This means irrespective of specific knowledge of the eventual context of use, or foresight into what such uses will be, GPAI developers should have certain due diligence requirements that they fulfill prior to release. This means that “opt-in” and “opt-out” approaches are inherently insufficient in mitigating risks and fostering informed decision-making. The data and design choices are usually shrouded in secrecy and a data subject or consumer cannot reasonably be expected to be aware of this process before making the choice to opt-in or opt-out of a specific end-use.
- **A legal disclaimer approach will effectively pass on all liability to downstream actors that lack sufficient resources, access, and ability to scrutinize and mitigate risks from GPAI systems.** It is worth noting that, at present, developers of GPAI include many of the largest and most well-resourced companies, compared to downstream actors that might be across a much larger spectrum. Such downstream actors should certainly be accountable for the specific context in which these models are applied (see Point 3) but to make them wholly liable for risks that emanate from data and design choices made at the stage of original development would be entirely misplaced. Furthermore, it would be unreasonable for upstream GPAI model creators to claim immunity for design and collection processes that only they exercise discretion and control over.

5. Regulation should avoid narrow methods of evaluation and scrutiny for GPAI that could result in a superficial checkbox exercise. This is an active and hotly contested area of research and should be subject to wide consultation, including with civil society, researchers and other non-industry participants.

GPAI systems must go through rigorous diligence, validation, and scrutiny before they are deployed or released publicly. Recent proposals that would include GPAI models within the scope of the AI Act either delegate the details of specific requirements to the future (to be specified by the Commission) or seek to already specify them within the text of the AI Act. Regardless of when and how these requirements are specified, lawmakers should note that there is no stable "state of the art" standard of evaluation that exists today which captures the range of risks and harms inherent to GPAI. It is also crucial to ensure wide ranging public engagement on any future specification of standards and requirements, including with civil society, researchers and other non-industry participants. However, it is important that for any regulatory approach to consider the following:

- **Regulation should leave adequate scope for broad and robust scrutiny at the original developer stage:** Standardized documentation practice and other approaches to evaluate GPAI models, specifically generative AI models, across many kinds of harm are an active area of research. Such practices are also necessary to ensure effective enforcement of AI regulations: without such documentation practices regulators will essentially be left in the dark when evaluating such systems. Documentation can provide structure to what developers should be assessing and making transparent that builds off of existing proposals such as Data Statements, Data Sheets, and Model Cards. However, these artifacts should be only one step in a broader evaluation process. Evaluation suites for the base model require significant investment and development across many dimensions, and should be easily accessible and runnable for developers. For example, evaluating harmful biases in a base model should cover many protected individual characteristics, types of biases such as stereotypes, and will differ by local and national context and application.
- **Regulation should not endorse narrow benchmarks for evaluation:** Different GPAI systems have different levels of maturity in benchmarks for evaluation. These benchmarks are typically oriented around performance metrics, such as accuracy, of such models against some held out evaluation dataset. However, there are multiple issues that arise in this method. For more mature benchmarks for certain classes of GPAI (e.g., facial recognition), while accuracy may be high across the evaluation dataset, it may perform much worse for racial and gender minorities. This has been the subject of vigorous academic debate in the domain of algorithmic fairness. Secondly, for very new forms of GPAI (e.g. language generation with LLMs), there are very few agreed-upon evaluation strategies and metrics. New technologies and the institutions which produce them move the goalposts with the development of new models. The lack of scientific consensus on evaluation exacerbates the problem of fairness in these systems. Lastly, the process of benchmarking may not be well-suited for mitigating risks in GPAI. Because tasks are too general, it is difficult to ensure that GPAI works for all possible use cases. Evaluation strategies also need to be attentive to the particularity of the task and how it may propagate information harm. For instance, whether a model is made widely available or is developed and used for a small community can influence whether potential harms are distributed widely or narrowly within a consenting community. For such evaluations to be more effective, more information sharing along the AI value chain would also be beneficial.

**For further information contact:**

Dr. Timnit Gebru, Founder and Executive Director, Distributed AI Research Institute (timnit@dair-institute.org)

Dr. Alex Hanna, Research Director, Distributed AI Research Institute (alex@dair-institute.org)

Amba Kak, Executive Director, AI Now Institute; Senior Research Scholar at Northeastern Khoury College of Computer Science (amba@ainowinstitute.org)

Dr. Sarah Myers West, Managing Director, AI Now Institute (sarah@ainowinstitute.org)

Maximilian Gahntz, Senior Policy Researcher, Mozilla Foundation (max@mozillafoundation.org)

Irene Solaiman, Policy Director, Hugging Face (irene@huggingface.co)

Dr. Mehtab Khan, Associate Research Scholar at Yale Information Society Project, (mehtab.khan@yale.edu)

Dr. Zeerak Talat, Independent researcher (Computer science and artificial intelligence), (z@zeerak.org)

## Signed:

Dr. Timnit Gebru, Founder and Executive Director, Distributed AI Research Institute

Dr. Alex Hanna, Research Director, Distributed AI Research Institute

Amba Kak, Executive Director, AI Now Institute

Dr. Sarah Myers West, Managing Director, AI Now Institute

Irene Solaiman, Policy Director, Hugging Face

Dr. Mehtab Khan, Associate Research Scholar at Yale Information Society Project,

Dr. Zeerak Talat, Independent researcher (Computer science and artificial intelligence),

Maximilian Gahntz, Senior Policy Researcher, Mozilla Foundation

Mark Surman, President and Executive Director, Mozilla Foundation

Dr. Abeba Birhane, Senior Fellow in Trustworthy AI, Mozilla Foundation

Meredith Whittaker, Chief Advisor, AI Now Institute

Frank Pasquale, Brooklyn Law School

Arvind Narayanan, Princeton University

Sayash Kapoor, Princeton University

Matthias Spielkamp, AlgorithmWatch

Angela Müller, AlgorithmWatch

Dr. Reuben Binns, University of Oxford

Suresh Venkatasubramanian, The Center for Technological Responsibility, Brown University

Sorelle Friedler, Haverford College

Janet Haven, Data & Society Research Institute

Jill Jung, Data & Society

Brandie Nonnecke, UC Berkeley

Os Keyes, University of Washington

Cedric Whitney, UC Berkeley School of Information

Ali Alkhatib, Center for Applied Data Ethics

Dr. Jennifer King, Stanford Institute for Human-Centered Artificial Intelligence

Luc Rocher, Oxford Internet Institute

Dr. Niels ten Oever, Assistant Professor in AI and European Democracies and co-Principal Investigator Critical Infrastructure Lab, University of Amsterdam

Dr. Ben Green, University of Michigan

Dr. Roel Dobbe, Delft University of Technology

Sacha Alanoca, Harvard Kennedy School

Leon Derczynski, IT University of Copenhagen

Olle Häggström, Chalmers University of Technology

Markus Anderljung, Centre for the Governance of AI

Jonas Schuett, Centre for the Governance of AI

Emma Bluemke, Centre for the Governance of AI

Michael Aird, Rethink Priorities

Noemi Dreksler, Centre for the Governance of AI

Sean O hEigeartaigh, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Kerry McInerney, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Haydn Belfield, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Dr. Stephen Cave, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Giulio Corsi, Leverhulme Center for the Future of Intelligence, University of Cambridge

Anna Katariina Wisakanto, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Jeffrey Gleason, Northeastern University

Lennart Heim, Centre for the Governance of AI

Jai Vipra, Centre for the Governance of AI

Jessica Newman, UC Berkeley AI Policy Hub

Ramak Molavi, The Law Technologist

Dr. Daniel Leufer, Access Now

Nico Mialhe, The Future Society

Dr. Dan Hendrycks, Center for AI Safety

Raja Chatila, Sorbonne University, Paris

Cyrus Hodes, AIGC Chain

Philippe Huberdeau, Scale-UP Europe

Simon Mueller Stansbury, The Future Society,

Samuel Curtis, The Future Society

Dr. Anthony M. Barrett, UC Berkeley AI Security Initiative

David Kreuger, University of Cambridge

Niki Iliadis, The Future Society

Jess Whittlestone, Centre for Long-Term Resilience

Tegan Maharaj, University of Toronto

Ben Winters, EPIC

## Institutional Signatories

AI Now Institute

Distributed AI Research Institute

Mozilla Foundation

AlgorithmWatch

Data & Society Research Institute

Center for Technological Responsibility, Brown University

Center for Applied Data Ethics, University of San Francisco

AI & Emerging Tech Caucus, Harvard Kennedy School

Aspiration Tech