Comments on the Final Trilogue Version of the AI Act

Philipp Hacker[1]

This version: January 23, 2024

partially drawing on my Statement on the AI Act Trilogue Results of Dec. 9, 2023, and on the takeaways from the AI House Davos Panel on International AI Regulation of January 16, 2024

Contents:

[1] Professor Dr. Philipp Hacker, LL.M. (Yale), Research Chair for Law and Ethics of the Digital Society, European New School of Digital Studies, European University Viadrina Frankfurt (Oder); co-lead, RECSAI.

I. General Observations

The final text is a political compromise and far from perfect. But, overall, I think it is better to have it than not to have it, for three reasons.

- First, minimum rules are established for foundation models, taking significant risks to public safety into account.
- Second, we now have minimum standards for protecting privacy and personal data as well as political freedoms in the context of remote biometric identification (RBI). In both areas, foundation models and RBI, more safeguards would be possible and desirable, but it is still better than nothing.
- Third, it will be possible for companies to come up with codes of conduct that can then be endowed with general validity by the Commission – regulated self-regulation. This provides flexibility, room for concrete sectorial implementations, and adds industry expertise.

The main points of critique are also three, in my view.

- First, the alignment with existing sectorial regulation is quite incomplete. This adds unnecessary and highly detrimental red tape.

- Second, compliance costs will be substantial. This is not a problem for foundation model providers, who need to invest massively to train a model, anyways – but for SMEs developing more narrow AI models.
- In RBI, we need, at least, a European supervision and monitoring. Otherwise, in countries with severe democratic backsliding, national agencies will control the national police. This is a recipe for disaster when it comes to facial recognition in public spaces, targeted at democratic opponents.

In the following, I offer some more detailed thoughts.

II. AI Definition

Revised OECD Definition, as incorporated into Article 3(1) AI Act: "An AI system is a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."

In this unqualified form, this is a definition of software, not of AI. Take an auto-sum function in an excel sheet. It has an objective (building a sum), input (entries), and an output that may influence environments (as per the relevance of the sum for any decisions). So, the only distinguishing criterion for AI on this definition is "infers".

Recital 6 now clarifies this to a certain extent: "The notion of AI […] should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations. A key characteristic of AI systems is their capability to infer. […] The techniques that enable inference while building an AI system include machine learning approaches that learn from data how to achieve certain objectives; and logic- and knowledge-based approaches that infer from encoded knowledge or symbolic representation of the task to be solved. The capacity of an AI system to infer goes beyond basic data processing, enable[ing] learning, reasoning or modelling."

So the two main approaches to AI, machine learning on the one hand and symbolic or knowledge-based approaches on the other, are now packaged into the term "inference". This is reassuring, but would still leave some loopholes, particularly concerning simple rule-based systems and statistical modeling. The Recitals do not reduce legal uncertainty to 0, but at least ensure that simple software, such as the auto-sum function in Excel, does not qualify as AI. one may either say that the auto sum function was defined solely by natural persons or that it involves basic data processing. While there is no clear red line, examples in guidelines would help enormously.

Against this background, the term autonomy is not decisive anymore. It is described in Recital 6 as "some degree of independence of actions from human involvement and of capabilities to operate without human intervention". The Recital should have made it clear that it takes 1) independence from human intervention (=automation) AND 2) significant adaptability/learning capacity (before or after deployment) to qualify for "AI-relevant inference". Since the latter is

only hinted at obliquely, the burden of separating simple software from AI now rests fully with the term "inference".

III. Foundation Models/general-purpose AI systems

1. What's new?

The final version of the AI Act introduces two changes in the regulation of foundation models (now called general-purpose AI systems again) vis-à-vis the political compromise text that became known in December, particularly regarding transition periods and copyright considerations.

- **Transition periods:** Foundation models are granted a 12-month period to comply with the AI Act rules from the date of their application. However, a new provision extends this transition period to 24 months for foundation models that are already on the market at the moment these AI Act rules come into effect (Art. 83(3)).
- **Copyright**: The AI Office is now tasked with providing a template for the detailed summary of the content used for training foundation models, facilitating transparency and compliance. The summary should list "the main data collections or sets that went into training the model, such as large private or public databases or data archives, and [provide] a narrative explanation about other data sources used" (Recital 60k).

2. Critique of GPAI Regulation for all Foundation Models

The following provisions are missing from the minimum requirements for all FM providers:

- **Cybersecurity**: It is imperative to ensure an adequate level of cybersecurity protection for all foundation models, particularly given the current geopolitical situation with increasing threats and wars. Insufficient cybersecurity measures propagate down the AI value chain and may open backdoors to a wide variety of applications for malicious actors, with a state or non-state background. This is a very serious threat, and strategic rivals are already actively exploiting it for industry espionage and military sabotage.

- **Content Moderation**: Expanded content moderation is necessary to mitigate the proliferation of hate speech and fake news. The lack of rules concerning robust content moderation measures is concerning. Experiments have shown that foundation models are prone to provide illegal outputs, including hate speech. Making sure that this potential cannot be abused by malicious actors seems paramount. Furthermore, content moderation must ensure that foundation models behave appropriately particularly also in dealing with "advice" provided by foundation models concerning physical or mental health problems. Overall, if the FLOPs threshold defining systemic risk models remains at 10^25, most powerful GPAI models will only have to meet the minimum standards. This oversight can be rectified by mandating, for all FMs, a compliance system to prevent illegal outputs, feasible for companies of varying scales, including Aleph Alpha and Mistral. The compliance system should ensure, via state-of-the-art technical and organizational measures, that content generated by AI, whether audio, image, video, or text, abides by the laws of Member States from which the model is accessible.
    - **Extension of DSA Provisions**: The provisions of Articles 16 and following of the Digital Services Act, including trusted flaggers and a notice-and-

action mechanism, should urgently be extended to the domain of Generative AI (see our paper https://dl.acm.org/doi/abs/10.1145/3593013.3594067). The reason for this is to establish a more effective and decentralized system for flagging and removing illegal content generated by AI systems to stem the tide of hate speech and hallucinations still plaguing GenAI – crucial ahead of the next global election cycles (US, EU, and beyond). This mechanism would bolster the existing content moderation framework by incorporating community-driven oversight. It would ensure a broader base for monitoring and mandate a quick response to violations highlighted by trusted flaggers (e.g., registered NGOs).

- **AI Safety**: There is an urgent need for comprehensive strategies to mitigate risks associated with cyber malware and biochemical terrorism. Again, even smaller foundation models may pose significant risks here. Mandatory provisions for all FM providers are essential. This includes
    - Mandatory red teaming for all FMs. This is basic industry practice.
- **Energy Consumption**: Providers must track, document, and report on the known or estimated energy consumption of the model.
    - This, at least, is a sensible measure.

3. Access for Vetted Researchers:

- Vetted researchers should have the right to access foundation models, akin to Article 40 of the Digital Services Act (DSA). The rationale for this is to allow for independent verification of stress tests and benchmarks, as well as decentralized monitoring. It's great that companies are doing research on this, but all results must be verified externally – that's just standard academic practice. Such access ensures that oversight does not solely rest with the providers of the models alone (and notoriously resource-constrained regulatory bodies) but involves the academic community at large.

4. Systemic-Risk GPAIS

Additional obligations for systemic-risk FMs have been rightfully introduced:

- **Evaluation, including Red Teaming**: Providers must perform model evaluations using state-of-the- art protocols and tools. There is a necessity to conduct and document adversarial testing (red teaming) to identify and mitigate systemic risks.
- **Risk Assessment and Mitigation:** More generally, systemic risks at the union novel must be assessed and mitigated via a comprehensive risk management system. This is primarily geared towards the prevention of major accidents, disruptions of critical sectors and serious consequences to public health and safety; any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; and the dissemination of illegal, false, or discriminatory content (Recital 60m).
- **Cybersecurity**: Maintaining an adequate level of cybersecurity for both the AI model and its physical infrastructure is non-negotiable.
- **Incident reporting**: to the AI Office

This is fairly comprehensive and good. But the threshold of 10^25 FLOPs for a default categorization of systemic risk models is too high. Currently, to my knowledge, only GPT-4, potentially Gemini, and perhaps one or two other models, surpass this threshold (see the

very useful study by The Future Society: https://thefuturesociety.org/wp-content/uploads/2023/12/EU-AI-Act-Compliance-Analysis.pdf).

Lowering this to 10^24 FLOPs would be more inclusive of large models that already demonstrate systemic risks and are currently on the market (GPT-3.5; Claude; Bard).

Critique of Systemic Risk FMs/GPAIS:

- **FLOPs Threshold**: The threshold of 10^25 FLOPs for a default categorization of systemic risk models is too high. Currently, to my knowledge, only GPT-4, potentially Gemini, and perhaps one or two other models, surpass this threshold (see the very useful study by The Future Society: https://thefuturesociety.org/wp-content/uploads/2023/12/EU-AI-Act-Compliance-Analysis.pdf).
    - **Important:** Lowering this to 10^24 FLOPs would be more inclusive of large models that already demonstrate systemic risks and are currently on the market (GPT-3.5; Claude; Bard). This is what the Commission should do via a Delegated Act.
- **External Red Teaming**: While the rules are sound, red teaming would benefit from the involvement of external entities to ensure an unbiased and comprehensive assessment.
- **Compliance Costs:**

IV. Open Source Exemptions

1. Functioning

The Trilogue exempts AI models made accessible under under free and open source (together: OS) licences from its scope (Article 2(5g)). This denotes models that are "openly shared and where users can freely access, use, modify and redistribute them or modified versions thereof" (Recital 60i) due to their presumed positive effects on research, innovation and competition.

By way of a reverse exemption, open-source models are covered by the AI Act in each of the following cases:

- integration into prohibited AI practices (Article 2(5g));
- integration into a high-risk AI system (Article 2(5g));
- scenarios leading to transparency provisions according to Article 52 (e.g., interacting with humans) (Article 2(5g));
- the copyright provisions for GPAIS, including the mandatory copyright compliance regime ensuring the respect of opt outs by rightholders (Article 52c(2));
- other rules for all GPAIS if parameters, including the weights, the information on the model architecture, and the information on model usage, are not made publicly available (Article 52c(2));
- rules for systemic-risk GPAIS (Article 52c(2)).

2. Critique of OS Exemptions

- **OS Models Threshold**: The current broad exemption for powerful OS GPAI models up until a 10^25 FLOPs threshold is questionable. A lower threshold, including minimum standards for 10^24 FLOPs OS text models, would ensure that such models are regulated appropriately.

- **OS Prohibition:** For quite highly-performing OS models, such as 10^24 FLOP models, one may even consider a restriction on open sourcing. These models should only be made available in an API or hosted access model. Studies show that safety layers can be easily removed once a model can be fully downloaded. These models are essentially dual-use goods, and cannot be made freely available for any use and modification for the general public. Otherwise, we may facilitate significant public safety threats, including cyber malware, and bio- and chemical terrorism. On the other hand, recent Stanford HAI research suggests that empirical evidence for marginal threat increases by open-source models vis-à-vis existing online resources is limited.

V. Remote biometric identification

I should mention at the outset that I am not really an expert on remote biometric identification (RBI) and law enforcement. Still, in my view, the real-time RBI provisions in the AI Act raise four key points of critique (the first three concerning real-time RBI and the last one post RBI). It is important to stress, however, that they only constitute minimum protections which can be expanded by Member States and do not constitute a legal basis for RBI. Rather, such a legal basis must be independently provided for in Member State law, respecting the limits enshrined in the AI Act and Article 10 of the Law Enforcement Directive.

1. Concerns about the Independent Administrative Authority

The Act allows an independent administrative authority, rather than a judicial authority, to authorize the use of real-time RBI. This could be problematic as even formally independent agencies might be susceptible to government influence, as evidenced in Poland under PiS, Hungary and other countries. The reliance on administrative rather than judicial oversight may not provide sufficient checks and balances.

2. Enforcement by Member State Agencies

The AI Act delegates the enforcement of these rules to member state agencies, with the legal basis rooted in member state law. This approach is concerning, especially for member states with questionable records in rule of law. For instance, entrusting a Hungarian agency with the oversight of Hungarian police use of RBI against political opponents of the Orban government will be problematic. This is exacerbated in situations where pre-authorization of RBI use can be bypassed, allowing authorization post-facto within 24 hours in urgent cases. Such flexibility could lead to misuse, especially in countries where democratic institutions are weakening (democratic backsliding).

3. Risk of Function Creep in Surveillance

The implementation of RBI as part of surveillance architecture inherently carries the risk of 'function creep', where the technology may gradually be used for purposes beyond its original intent. This poses a significant threat to civil liberties and personal privacy, as the scope of surveillance could expand over time without adequate safeguards and oversight, as evidenced

in the recent CJEU Case C470/21, La Quadrature du Net and Others (https://curia.europa.eu/jcms/upload/docs/application/pdf/2023-09/cp230151en.pdf).

4. Post RBI

The post-collection provisions for RBI (post RBI) in the AI Act are governed by three main guardrails (Recital 58e): a general reference to the proportionality principle, the prohibition of indiscriminate surveillance, and the prohibition of circumventing the rules for real-time RBI.

Those limits are relatively weak. Hence, Member States will have to do the heavy lifting concerning post RBI. This delegation of responsibility to Member States is a significant point of concern, given that some EU Member States have demonstrated a propensity to disregard rule of law principles. The reliance on Member States for the enforcement of these broad and somewhat vague guardrails could lead to inconsistent application and potential misuse of RBI technology, particularly in jurisdictions where adherence to the rule of law is less stringent.

Importantly, this critique **should not lead to wholesale rejection of the AI Act**. Without the AI Act, even though those minimum guarantees contained in it would be missing - which would arguably be even worse from a sensible privacy perspective.

VI. Copyright

The copyright provisions within the AI Act are reasonably constructed.

1. Compliance regime and training content summary

- **Copyright Compliance Regime**: Providers must implement a policy that respects Union copyright law, utilizing state-of-the-art technologies where appropriate. This is tantamount to a compliance regime, putting in place organizational and technical measures to ensure heating the opt-out rights of rightholders. This makes sense as only companies systematically violating copyright provisions would not install such a system.
- **Training Content Summary**: Providers are also required to draw up a sufficiently detailed summary of the content used for training the AI model. As mentioned above (III.1.), the AI Office will now be tasked with providing a template for the summary.

2. Critique of Copyright Provisions

- **Detail of Summary**: It must be made explicit that the summary does not need to delve into individual training data points, which would be prohibitively expensive. Indeed, this is now been clarified in Recital 60k: The summary only needs to list "the main data collections or sets that went into training the model, such as large private or public databases or data archives, and [provide] a narrative explanation about other data sources used." This clarification is highly welcome.

VII. High-Risk Classification

Article 6(3) contains the "extra layer." It outlines the criteria for not classifying AI systems as high risk despite used in one of the Annex III sectors if and when there is no significant risk to fundamental rights. The cases exempt from high-risk classification are:

- Narrow procedural tasks;
- Mere preparatory tasks;
- Improvements to the results of previously completed human activities;
- Decision or outlier detection without replacing human assessments.

These criteria provide a relatively clear framework for determining whether an AI system qualifies as high risk. It follows from a risk-based approach that high-risk regulation is indeed unnecessary if the AI system, in the specific use case, does not carry significant risks, for example, if it schedules appointments for Annex III activities (e.g., job talks). While this does provide companies with a way out of the AI Act, the risk of misclassification is significant, which should act as a deterrent to companies simply claiming not to pose significant risks with their AI systems. Overall, in my view, this constitutes a balanced and necessary provision within a risk-based framework.

VIII. Innovation

In my view, sandboxes will not be decisive. Successful startups do not need them right now and will likely not really need them later on. Rather, next to legal certainty, two things are crucial:

1. Alignment with existing sectoral regulation

The alignment with existing sectoral regulation stands out as particularly significant. This alignment ensures that the AI Act does not operate in isolation but rather complements and integrates with the broader regulatory framework. The AI Act must not leave critical areas doubly, inefficiently, or even conflictingly regulated (such as medical AI, credit scoring, and insurance). Rather, compliance with specific sectorial regulation should lead to a presumption of compliance with comparable rules in the AI Act. Unfortunately, this is only imperfectly realized.

According to Recital 54, "Providers of high-risk AI systems that are subject to obligations regarding quality management systems under relevant sectorial Union law should have the possibility to include the elements of the quality management system provided for in this Regulation as part of the existing quality management system provided for in that other sectorial Union legislation." This is a good start, implemented in Article 9(9) and Article 17(2a). However, it would be important to give significantly beyond this and to **develop an exact mapping between existing rules and those rules of the AI Act which are presumably complied with if the sectorial rules are fulfilled**. For some work in this direction concerning financial services, see https://taktile.com/articles/the-future-of-credit-underwriting-under-ai-regulation-implications-for-the-eu-and-beyond.

2. AI Value chain

Importantly, in the AI value chain, deployers assume the responsibilities of providers if they make substantial modifications to the model (Art. 28(1)(b)). This is of particular relevance if foundation models are used downstream and adapted to specific use cases. It remains clear that,

generally, deployers should not assume the responsibilities of providers, with liability and compliance attached, if they merely fine-tune a general-purpose model. Otherwise, this would send a chilling effect down the AI value chain, particularly for less technically savvy companies.

From a legal perspective, the key challenge lies in precisely defining the threshold for what constitutes a "significant modification," making it measurable and discernible. Striking a balance between fostering innovation and ensuring legal clarity becomes crucial in delineating the boundary where a modified model triggers new obligations.

To operationalize these principles, I suggest introducing a new recital and articles. Firstly, a new recital should emphasize that the obligations under Article 52c and d of the AI Act generally apply to the original GPAI providers who release a model. Secondly, a legal presumption is recommended, stating that further adaptations of the FM, excluding significant changes in risk profile, do not invoke new obligations. A new article, Art. 28(x) could embody this presumption, asserting that modifications to the originally released FM do not make the modifying entity a new provider, unless the changes significantly alter the model's risk profile. A corresponding recital could clarify that this presumption covers various modification techniques, such as pre-training, fine-tuning on additional data sets, knowledge distillation, and quantization. This presumption should not apply if the risk profile of the model changes very significantly, for example in cases of fine-tuning on a data set that is known to contain a significant amount of hate speech or of explicit removal of the safety layer of models.

IX. Regulated Self-Regulation and Safe Harbors

Code of Practice: The Commission’s power to approve codes of practice for GPAI will endow the Code with general validity within the Union. This is an excellent development. It presents an opportunity to leverage decentralized industry and expert knowledge and operationalize vague concepts for specific sectors. This helps to establish safe harbors for companies. These will be crucial to attract and retain companies, and talent, in the EU.

X. Employment

The discussion surrounding the legal basis of the finalized AI Act, particularly in relation to employment and worker protection, presents a challenging legal conundrum. The crux of the issue lies in Article 114(2) of the Treaty on the Functioning of the European Union (TFEU) as the specific prohibition for regulating aspects of worker rights and interests via Art. 114 - which is, however, the legal basis of AI Act. Article 114 TFEU is predominantly concerned with harmonizing legislative measures pertinent to the internal market. Its primary objective is to facilitate the smooth functioning of the market across the EU. The contentious point in the context of the AI Act is whether this Article constitutes an appropriate and sufficient legal basis for enacting regulations that significantly influence worker rights, an area traditionally governed by the social policies of individual Member States, and directives based on Art. 153 TFEU.

From a systematic perspective, **the rules relating to AI in employment may well be unconstitutional and open to a legal challenge before the CJEU**.

This concern is not unique to the AI Act. Similar legal questions have been raised in relation to the Digital Markets Act (DMA) and its basis in Art. 114 TFEU, see, e.g., https://awards.concurrences.com/en/awards/2022/academic-articles/why-the-proposed-dma-might-be-illegal-under-article-114-tfeu-and-how-to-fix-it.

XI. Fundamental rights impact assessment

The fundamental rights impact assessment (FRIA) for high-risk AI systems according to Article 29a comes with the best intentions, but will likely remain a paper tiger, generating efforts and costs without much effect. First, it remains unclear to what extent this impact assessment will go beyond the general risk assessment according to Article 9, which also includes mapping and mitigating risks to health, safety and fundamental rights. Second, doctrinally, it continues to be perplexing that private companies should even be subject to fundamental rights (unless specific situations arise in which the CJEU has indeed conveyed a direct horizontal effect on some Charter rights[2]). This is likely the reason why Article 29a(1) now restricts the ambit of the FRIA to deployers covered by public law and providing public services.

However, banks, insurance companies, as well as those companies active in the fields of education, healthcare, or housing, for example, are still supposed to be covered, raising the complex question of how these companies should, from a strictly legal perspective, be in a position to violate fundamental rights in the first place, except for the right to non-discrimination for which the CJEU has broadly argued that it does apply between private parties.[3]

Finally, it is foreseen that the exercise will be completed through a questionnaire including an automated tool – which, in the end, **will likely devolve into a box-checking exercise without much effect.**

XII. Right to an explanation

The introduction of a right to an explanation in the AI Act under Article 68c, while a step forward, has its limitations when scrutinized against the broader discourse on explainable AI within the academic community.

The provision applies specifically to high-risk AI systems, merging this requirement with the elements of Article 22 GDPR, and granting affected individuals the right to understand the role of the AI system in the decision-making process and the main elements of the decision. However, this does not, at least not necessarily, encompass a comprehensive explanation of the AI system's inner workings, which is a central topic in the field of explainable AI.

Key concepts in explainable AI, such as feature salience (identifying which features of the input data were most influential in the decision-making process) or counterfactual explanations (understanding how a slight change in input could have led to a different decision), are not covered by this provision (except in rare cases where this could be construed to be part of the main elements of the decision). This limitation indicates a gap between the legal framework's

[2] This concerns primarily Art. 21(1) und Art. 31(2) of the Charter, see CJEU, Case C-414/16 (Egenberger); C-68/17 (IR); Joint Cases C-569/16 and C-570/16 (Bauer und Willmeroth).
[3] See n. 2.

approach to AI transparency and the more technical, detailed explanations discussed in XAI scholarship. On the other hand, it does not tie developers to specific explanation techniques, which is sensible giving the rapidly evolving nature of the field.

Adding to the critique of Article 68c AI Act, its practical utility is likely to be quite limited, especially when considered in light of recent jurisprudence that has strengthened the right to explanation under the General Data Protection Regulation (GDPR).

Notably, the SCHUFA cases (CJEU) and the rulings in the Uber and Ola cases by the Amsterdam Court of Appeals have significantly expanded the scope of the right to an explanation under Article 15(1)(h) of the GDPR. These cases have set precedents that emphasize the importance of transparency and the right of individuals to understand how decisions affecting them are made, especially when these decisions are derived from automated processing. **It is unlikely that Article 68c AI Act is going to play a major role besides this more far-reaching jurisprudence**, particularly if the Uber and Ola judgments are ultimately confirmed.

XIII. Tight timeline

The tight timeline for the adoption of the EU's AI Act is challenging but a necessary measure. The pressing need to finalize this legislation before the upcoming EU elections has imposed a tight schedule. While this accelerates the legislative process, it is crucial to ensure that the regulatory framework for AI is in place in a timely manner. This urgency, while it may strain the thoroughness and deliberation typically desired in legislative processes, is a pragmatic response to the political and technological realities of the moment. It underscores the importance the EU places on establishing a robust legal framework for AI ahead of significant political events. Despite the challenges this timeline presents, it reflects a commitment to swiftly addressing the rapidly evolving landscape of AI technologies and their implications.

This urgency, while understandable from a political standpoint, has led to compromises in the legal package. Sometimes, quality was sacrificed to speed. A notable area of not fully convincing compromise is the definition of AI (the definition of autonomy is vague and too broad; it should also include some learning capability) and the rules regarding foundation models (too permissive, disregarding much AI safety research). The rush to complete the legislation could potentially compromise its quality in these and other areas. Despite these concerns, the act is still likely to pass. **The possibility of France and Germany forming a blocking minority seems unlikely, though it remains a scenario to consider**. But Germany, in particular, would have to spend a lot of political capital and would rightly be perceived as a destructive force in one of the key legislative areas of our time, with massive implications for the economy and safety of the EU and beyond.

Going forward, the timeline for the applicability of the AI Act is also very tight (see below, Transition Periods).

XIV. Compliance preparation for businesses

For businesses, preparation for the AI Act's implementation is critical. The first step is to review all current AI systems to determine their classification under the act, particularly identifying

any high-risk applications. Companies should also examine the transparency requirements as per Article 52. Moreover, the need to set up compliance regimes especially for AI systems that may fall under the category of foundation models, now termed general-purpose AI systems. importantly, while the AI Act requirements mainly address the original developers, making significant changes to the model may bring deployers under the ambit of the foundation model rules. It's important to note that compliance requirements for general-purpose AI systems are not limited to the AI Act but extend to the General Data Protection Regulation (GDPR), non-discrimination law, and copyright laws. Our recent article available at https://arxiv.org/abs/2401.07348 provides further insights into this.

Additionally, businesses should stay informed about technical standards emerging from organizations like ISO or CEN-CENELEC. These standards will be pivotal in bridging the gap between legal requirements and technical machine learning procedures. Finally, developing a code of conduct tailored to the organization's AI applications can facilitate smoother implementation of the AI Act's requirements. This proactive approach will help ensure compliance and minimize disruption to business operations.

XV. Going Forward

1. Changing the Narrative in AI Regulation and Governance

In the discourse of AI governance, there's an urgent need to shift the narrative from a risk-centric to an opportunity-centric viewpoint. High-risk AI applications often present significant opportunities, especially in sectors like healthcare, education, and employment. These areas are ripe for transformative impacts through AI, provided the implementation is done thoughtfully and responsibly. This perspective emphasizes the potential of AI to bring about positive, high-impact changes, while not undermining the importance of cautious and well-regulated deployment.

2. Transition Periods in the AI Act

The AI Act introduces a structured timeline for compliance, catering to different categories of AI systems (Art. 85):

- Prohibited practices have a compliance period of 6 months.
- Foundation models are given 12 months; 24 months if already on the market at this moment (Art. 83(3)).
- Annex III high-risk systems are allocated 24 months; systems on the market at this moment are not covered at all unless they undergo significant changes in the design after those 24 months (Art. 83(2)).
- Annex II high-risk systems have a 36-month period; systems on the market after 24 months are not covered at all unless they undergo significant changes in the design after those 24 months (Art. 83(2)).

This staggered approach allows for a gradual adaptation to the regulatory environment. However, organizations must act promptly to prepare and align their AI strategies with these regulatory timelines.

Importantly, the exemption of existing high-risk systems opens a significant loophole and contradicts general legislative practices in the field of product safety. After all, if a risk

exists, it is irrelevant if the product was on the market at the time of the applicability of the Act or if it is only introduced after that moment. The transition period of 24/36 months is also not so short that the manufacturers of existing systems would not be in a position to adapt to the AI Act.

3. Compliance Costs

The costs associated with compliance will be substantial, particularly affecting smaller organizations with limited resources. While well-funded companies may navigate these financial demands more easily, smaller enterprises might struggle. This disparity raises real concerns about maintaining a competitive and diverse AI ecosystem, where smaller players can also thrive and innovate.

4. Standards and Guidelines

Standards and guidelines will play a crucial role in the practical implementation of AI governance. Despite valiant efforts by organizations like ISO and CEN-CENELEC, there remains a challenge in achieving clarity, especially in sector-specific contexts. This lack of crystal-clear guidelines could hinder the effective and consistent application of standards across different AI applications.

5. International Collaboration

The increasing necessity for international collaboration in AI regulation and the sharing of best practices cannot be overstated. Such collaborative efforts are essential in creating a globally coherent and effective AI governance framework, addressing the transnational nature of AI technologies and their impacts.

6. Geopolitical Stakes

The geopolitical stakes in AI governance are immense and continually escalating. The way AI is regulated and deployed has significant implications for global power dynamics, economic competitiveness, and national security. This aspect underscores the need for a strategic approach to AI governance that considers these broader geopolitical implications.

7. Race for Safe Harbors

There is an ongoing competition among nations to establish themselves as safe harbors for AI companies. This race is driven by the desire to attract AI innovation and investment, positioning these countries as leaders in the emerging AI-driven economy. This dynamic has significant implications for global AI development and governance.

8. Soft Law Frameworks

Soft law frameworks, such as the NIST Risk Management Framework, are gaining prominence in AI implementation. The work by NIST and NAIAC in this domain is commendable, providing flexible and adaptive guidelines that complement hard law regulations. These

frameworks are pivotal in guiding AI practices, especially in areas where formal legislation is still evolving.

9. Testing Tools and Open-Source in AI

The development of testing tools, particularly for white box testing, is crucial in the AI landscape. Open-source models are leading the charge in this area, offering transparency and ease of testing. However, the open-source approach presents a complex challenge in regulating foundation models. There's a delicate balance to strike between ensuring safety (preventing misuse by malicious actors), promoting competition, facilitating testing, and ensuring equal access. For cutting-edge AI models, like those with ChatGPT-like capabilities, safety concerns may be prioritized over other aspects. However, this recommendation may differ for less advanced models, where the risk profile is lower.