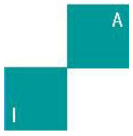
Inteligencia Artificial 27(73) (2024), 55-64
doi: 10.4114/intartif.vol27iss73pp55-64



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

The EU AI Act: A pioneering effort to regulate frontier AI?

Guillem Bas[1,A], Claudette Salinas[1], Roberto Tinoco[1], Jaime Sevilla[1]
[1] ORCG – Observatorio de Riesgos Catastróficos Globales
[A] gbasg@riesgoscatastroficosglobales.com

Abstract The emergence of increasingly capable artificial intelligence (AI) systems has raised concerns about the potential extreme risks associated with them. The issue has drawn substantial attention in academic literature and compelled legislators of regulatory frameworks like the European Union AI Act (AIA) to readapt them to the new paradigm. This paper examines whether the European Parliament's draft of the AIA constitutes an appropriate approach to address the risks derived from frontier models. In particular, we discuss whether the AIA reflects the policy needs diagnosed by recent literature and determine if the requirements falling on providers of foundation models are appropriate, sufficient, and durable. We find that the provisions are generally adequate, but insufficiently defined in some areas and lacking in others. Finally, the AIA is characterized as an evolving framework whose durability will depend on the institutions' ability to adapt to future progress.

Resumen La aparición de sistemas de inteligencia artificial (IA) cada vez más capaces ha generado preocupación sobre los posibles riesgos extremos asociados con ellos. La cuestión ha atraído una atención sustancial en la literatura académica y ha obligado a los legisladores de marcos regulatorios como el Reglamento de IA de la Unión Europea (RIA) a readaptarlos al nuevo paradigma. Este artículo examina si el borrador del RIA del Parlamento Europeo constituye un enfoque apropiado para abordar los riesgos derivados de los modelos punteros de IA. En particular, discutimos si el RIA refleja las necesidades de políticas diagnosticadas por la literatura reciente y determinamos si los requisitos que recaen sobre los proveedores de modelos fundacionales son apropiados, suficientes y duraderos. Encontramos que las disposiciones son en general adecuadas, pero insuficientemente definidas en algunas áreas e inexistentes en otras. Finalmente, el RIA se caracteriza por ser un marco en evolución cuya durabilidad dependerá de la capacidad de las instituciones para adaptarse al progreso futuro.

Keywords: Artificial Intelligence (AI), frontier models, regulation, governance, European Union (EU)
Palabras clave: Inteligencia Artificial (IA), modelos punteros, regulación, gobernanza, Unión Europea (UE)

1 Introduction

Artificial intelligence (AI) has progressed significantly in the past years due to advances in machine learning [1] and the growth of the computing power used to train AI systems [2]. One of the most remarkable demonstrations of this evolution has been the creation of large models with unprecedented capabilities, such as mastering natural language. This paper refers to these emerging systems as *frontier models*.

Anderljung *et al.* [3] define *frontier models* as "highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety". For their part, Shevlane *et al.* [4] describe them as those models "close to, or exceeding, the average capabilities of most capable existing models, and different from other models, either in terms of scale, design, or their resulting mix of capabilities and behaviours". Several industry leaders have also focused their governance efforts on frontier models, which they define as "large-scale machine-learning models that exceed the capabilities currently present in the most advanced existing models, and can perform a wide variety of tasks" [5].

ISSN: 1137-3601 (print), 1988-3064 (on-line)

Frontier models hold the promise of immense societal impact, both positive and negative. On the one hand, current AI models could raise global GDP by 7% [6] and enable 134 out of 169 targets across all Sustainable Development Goals, by making education more accessible, supporting the provision of essential goods and services, or underpinning a more efficient use of resources, among others [7]. AI has also contributed to scientific breakthroughs such as solving the protein-folding problem [8], controlling tokamak plasmas [9], and discovering antibiotics against pathogens with multidrug resistance [10]. Current and future frontier models are only expected to untap an even greater potential.

Notwithstanding, the list of risks is also large and requires great care [11]. In that regard, frontier models already possess dangerous capabilities that can be exploited by malicious actors [12]. This includes threats such as the enhancement of cyber-offenses [13] [14], the manipulation of humans [15] [16], or the development of biological and chemical weapons [17] [18]. Even more worryingly, future AI systems could act autonomously and inflict unintended damage if their behavior is not aligned with human values and goals [19] [20] [21].

Accordingly, an increasing number of actors have pushed for the creation of regulatory frameworks that address the challenges posed by AI. One of the most significant initiatives is the European Union's Artificial Intelligence Act (AIA), initially proposed by the European Commission in 2021 [22]. At the date of publication of this paper, both the Council of the EU [23] and the European Parliament (EP) [24] have adopted their own positions on the regulation and are negotiating the definitive outcome. This paper focuses on the EP's draft, published in May 2023, because it is the most recent public version. As such, some of its proposed provisions have become a central topic of debate among EU institutions.

The relevance of this work is two-fold. First, through the so-called "Brussels effect", the AIA could diffuse to the rest of the world if the EU leverages its ability to influence global governance through its market power, which incentivizes both EU and non-EU companies to make EU-compliant products and preserve customers [25]. Second, as the first comprehensive regulatory framework on AI, the AIA could possess high information value, as it will constitute a relevant experience to learn from and possibly serve as an archetype to inspire future legislation elsewhere.

The rest of this paper is divided into five sections. First, we provide an overview of the main policies for frontier models proposed in academic literature. Second, we identify the main requirements set in the EP's draft for providers of foundation models. Third, we analyze how these obligations interact with the frameworks proposed in the literature. Fourth, we discuss how future-proof the AIA might be by considering the possibilities to update it. And fifth, we finish with a conclusion that wraps up the main findings.

2 Theoretical proposals to regulate frontier AI models

The proliferation of frontier models has been responded to by calls for greater governance efforts, which has translated into a rapidly growing body of literature. In this section, we provide an overview of the most discussed proposals, roughly divided into reporting mechanisms, assessment methods, and post-deployment control.

An important first step for effective governance is regulatory visibility, that is, identifying and addressing the appropriate regulatory targets by having a strong understanding of the technological ecosystem [3]. In this context, computational resources are often considered a useful node for AI governance. Since there is a strong correlation between the amount of compute used for training and the capabilities of the resulting model [26], detecting where large amounts of compute are being used may enable governments to develop early awareness of which actors are likely to be developing and deploying highly capable systems [27]. This could be possible if developers report their training resources or if regulators monitor training runs or even the semiconductor supply chain [28]. Ultimately, the most resource-intensive models could be subject to especially stringent policies such as audits, risk management systems, and post-deployment safeguards.

Auditing is "a structured process by which an organization's present or past behavior is assessed for consistency with relevant principles, regulations, or norms", and it has been used extensively in industries such as finance and air travel [29]. In order to avoid biases and conflicts of interests, experts usually recommend that these processes be carried out by a third party [30] or an internal audit team that is organizationally independent from senior management [31].

To adapt audits to AI, Mökander *et al.* [32] propose an approach that integrates three distinct layers: supervising the auditee's organizational procedures, incentive structures, and management systems; assessing the functionalities and impact of the products and services through which the model is deployed; and identifying the capabilities and limitations of the model itself. Due to the nature of the technology, the latter has received particular attention.

Model evaluations are empirical assessments of a model's properties, and are considered crucial to identify dangerous capabilities and the propensity of models to apply their capabilities for harm [4]. Even though the list of potential dangers is extensive, organizations working on evaluations have tended to focus on specific critical capabilities, such as autonomous replication and adaptation in the case of ARC Evals [33].

Another relevant method to assess an AI system is red teaming, which is "a structured effort to find flaws and vulnerabilities in a plan, organization, or technical system, often performed by dedicated 'red teams' that seek to adopt an attacker's mindset and methods" [29]. Some leading organizations have successfully carried out this type of exercise to test their models by adversarially probing it for harmful outputs and then updating the model to avoid such outputs [34] [35].

Besides model evaluations and red teaming exercises, Koessler and Schuett [36] suggest other potential techniques to be used for risk identification, such as scenario analyses, which use forward thinking to develop future scenarios, or the fishbone method, which uses backward reasoning from a risk to its sources. More exhaustively, Barrett *et al.* [37] provide an overview of interventions and translate some techniques and high-level principles into actionable guidance, which is essential to guarantee a comprehensive risk assessment.

During training, risk assessments could inform the need to adjust the training methods or carefully scale the model [4]. Similarly, risk assessment could condition how AI systems are deployed: the greater the risk, the greater the restrictions to be enforced, ranging from proceeding with certain caution to cancelling deployment altogether [3].

Even after deployment is approved, AI providers would still have to implement guardrails for several purposes. One of them would be preventing misuse, which could be done through measures such as restricting usage or inhibiting misuse-relevant tasks [38]. Another objective would be preventing users from circumventing a model's restrictions to modify or reproduce it. In that regard, Shevlane [39] proposes *structured access* as an emerging paradigm that constructs a controlled, arm's length interaction between an AI system and its user.

Besides, a model's risk profile may evolve post-deployment, as new dangerous capabilities could emerge through scaling [40] or if users find unanticipated ways to enhance it [41] and exploit it for harm [42]. In that sense, it is considered important to reassess the model regularly [3] and track post-deployment logs through audit trails [29]. Proposed solutions to this problem include mitigation measures such as continuous fine-tuning [43] and reinforcement learning from human feedback [44]. If these and other practices are deemed insufficient, it would be necessary to have the legal and technical ability to quickly roll back deployed models on short notice [3].

The list above is not exhaustive or undisputed, but it reflects some of the most approved policies. In a recent survey to leading experts from AI labs, academia, and civil society (n = 51), 98% of respondents favored the mentioned assessment policies (model audits and evaluations, red teaming, and pre-deployment risk assessments); 98% and 97% agreed on monitoring systems and conducting post-deployment evaluations, respectively; and 78% supported the pre-registration of large training runs [45].

3 The European Parliament's proposal for the AIA

The main risk framework of the AIA is not based on the capabilities of AI systems but on their applications. As such, for instance, high-risk AI systems are those "falling under one or more of the critical areas and use cases referred to in Annex III [...] [that also] pose a significant risk of harm to the health, safety or fundamental rights of natural persons". Nevertheless, the AIA has been reformulated to include categories that do not fit in that framework. The most important example of this has been the introduction of the term *foundation model*, which the European Parliament defines as an "AI model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks".

In that regard, the most relevant fragment of the latest version of the AIA is Article 28b, which lays down obligations for providers of foundation models. Article 28b(2) lists seven requirements while Article 28b(4) adds an additional three for those models characterized as generative AI. This paper only focuses on the former.

In summary, Article 28b(2) covers the following requirements: (a) identifying, reducing, and mitigating reasonably foreseeable risks; (b) implementing appropriate data governance measures; (c) achieving appropriate levels of performance, predictability, interpretability, corrigibility, safety, and cybersecurity assessed through appropriate methods such as model evaluation; (d) reducing energy use, resource use, and waste; (e) drawing up technical documentation and instruction for use; (f) establishing a quality management system; and (g) registering the model in an EU database.

As displayed in Table 1, the obligations laid down in Article 28b are mostly analogous to some of those falling on high-risk systems. Nevertheless, the connection is not explicit and the formulation in Article 28b tends to be

more generic and briefer. That is to say, the requirements listed there are expressed in high-level principles and condensed in one sentence per duty.

Table 1: Parallelisms between Article 28b and provisions for high-risk AI systems.

Article 28b(2)	**Provisions for high-risk AI systems**
Paragraph a)	Article 9 (Risk management system)
Paragraph b)	Article 10 (Data governance)
Paragraph c)	Article 9 (Risk management system), Article 13(1) (Interpretability), Article 14 (Human oversight), Article 15 (Accuracy, robustness, cybersecurity), Article 43 (Conformity assessment)
Paragraph d)	None
Paragraph e)	Article 11 (Technical documentation), Article 13(2) (Instructions)
Paragraph f)	Article 17 (Quality management system)
Paragraph g)	Article 60 (EU Database)

Other than that, Chapter 3 includes several provisions on post-market monitoring. Article 65(2) states that "where the national supervisory authority of a Member State has sufficient reasons to consider that an AI system presents a risk [...], it shall carry out an evaluation of the AI system concerned in respect of its compliance with [...] [the] Regulation". If the test is not approved, the authority "shall without delay require the relevant operator to take all appropriate corrective actions to bring the AI system into compliance, to withdraw the AI system from the market, or to recall it within a reasonable period, commensurate with the nature of the risk". According to Article 65(5), "where the operator of an AI system does not take adequate corrective action [...], the national supervisory authority shall take all appropriate provisional measures to prohibit or restrict the AI system's being made available on its national market or put into service, to withdraw the AI system from that market or to recall it". Additionally, as per the same article, "that authority shall immediately inform the Commission, the AI Office and the national supervisory authority of the other Member States without delay, of those measures".

Finally, Title VI covers institutions for governance. The European Parliament proposes, in Article 56, the establishment of the European Artificial Intelligence Office (the *'AI Office'*), an independent body of the Union with legal personality. This AI Office contrasts with the European Artificial Intelligence Board (the *'AI Board'*) previously proposed by the Commission and the Council, which was conceived as a coordination mechanism mostly limited to advising Member States and the Commission.

The AI Office would be empowered by a widened set of tasks listed in Article 56b, which includes some relevant assignments for the governance of foundation models. In particular, the AI Office would be responsible to "provide particular oversight and monitoring and institutionalize regular dialogue with the providers of foundation models about the compliance of foundation models [...] with Article 28b of [the] Regulation, and about industry best practices for self-governance" (Paragraph q). It would also have to "issue an annual report on the state of play in the development, proliferation, and use of foundation models alongside policy options to address risks and opportunities specific to foundation models" (Paragraph r).

4 Examining the AIA's suitability to regulate frontier models

As seen in the previous section, the EP's draft of the AIA covers some of the main policy needs diagnosed by recent literature. In this section, we discuss whether the provisions are appropriate and sufficient, based on a critical analysis of the terminology and policies proposed. In particular, the text identifies a need for more specific language and indicators in the definitions and highlights the importance of being more specific, unambiguous, and exhaustive with the obligations.

Initially, it is worth revisiting the scope of the AIA. As already mentioned, the latest version of the AIA mostly regulates unacceptable and high-risk AI systems, as categorized according to the original risk framework, and foundation models. The concept *foundation model* arguably includes all the recently released frontier models, but the EP's definition does not explicitly mention neither 'high capabilities' as a relevant variable nor objective indicators of these capabilities. However, the draft provides some ground for the definition of objective indicators that could help constitute other categories, such as *frontier models*. One of the most notable instances is the inclusion of the term *large training runs*, which refers to "the production process of powerful AI models that require computing resources above a very high threshold".

Some relevant actors within the EU ecosystem have supported the idea of reconsidering, once again, the terminology of the AIA. Zenner [46] has proposed a distinctive category called *systemic foundation models* to refer to "a small number of highly capable as well as systematically relevant foundation models", which would be operationalized through metrics such as the financial investment in the model, the amount of compute usage, or the scalability of deploying the model. Similarly, the Spanish presidency of the Council has proposed the introduction of a category of 'high-impact foundation models', defined as "any foundation model trained with large amount of data and with advanced complexity, capabilities and performance well above the average for foundation models, which can disseminate systemic risks along the value chain, regardless they are integrated or not in a high-risk system". In this case, the definition would also be specified by thresholds to be defined by the Commission [47].

With regards to policies, the AIA performs disparately at integrating reporting mechanisms, assessment methods, and post-deployment control. First, the EP's draft does not establish a clear reporting system, but it sets the ground for this system to exist. Providers of foundation models must register their models in a public database (Article 28b(2g)) and draw up extensive technical documentation (Article 28b(2e)) that includes, among others, a "description of the training resources used by the foundation model including computing power required, training time, and other relevant information related to the size and power of the model" (Annex VIII, Section C, Paragraph 7). For the reports to be effective, regulators would potentially need to ensure that the latter is not only available in the technical documentation but also easily accessible through the database.

Second, in the case of assessment measures, the AIA seems appropriate but not sufficient. Article 28b(2c) mentions "model evaluation with the involvement of independent experts" and "extensive testing during conceptualisation, design, and development", while Article 28b(2c) refers to "the identification, the reduction and mitigation of reasonably foreseeable risks". Notwithstanding, the Article appears to be non-specific and inconclusive. Model evaluation and testing, for example, are presented as examples of possible methods "to achieve [...] appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity", but they are not part of a mandatory conformity assessment as in the case of high-risk AI systems (Article 43). This lack of detail risks turning the obligations set in Article 28b into mere recommendations. Greater precision could be achieved by detailing the comprehensiveness of the referred practices and clarifying which ones should be compulsory.

In that regard, Recital 60h acknowledges that "given the nature of foundation models, expertise in conformity assessment is lacking and third-party auditing methods are still under development", so "the sector itself is [...] developing new ways to assess foundation models that fulfill in part the objective of auditing". Industry practitioners might indeed be the best suited to govern AI in the absence of dedicated multistakeholder institutions [48]. Nevertheless, leaving the definition of specific duties at the developers' discretion might elicit laxer procedures by some of these developers, as competitive pressures could incentivize companies to underinvest in precautionary measures [49]. Alternatively, the AIA could enact high-level obligations to be operationalized by later standards and fed back by practical experience. That is the case for conformity assessments of high-risk systems, which are already an important step to establish a Europe-wide ecosystem for conducting AI auditing [50].

Third, post-deployment policies seem to be the least covered by the AIA. Article 28b does not mention any besides the obligation of drawing up "intelligible instructions for use in order to enable the downstream providers to comply with their obligations". This contrasts with the obligations of deployers of high-risk AI systems, which include implementing human oversight (Article 29(1a)) or keeping the logs automatically generated by the AI system for ex-post audits of any reasonably foreseeable malfunction, incidents, or misuses (Article 29(5)).

By relying on compliance with the model's instructions, the AIA misses two important particularities of foundation models. On the one hand, large-scale and rapid releases facilitate access to a wide group of users, increasing the potential for harmful use cases [51]. On the other hand, the model could be substantially modified by adversarial actors [52]. As for the former, the AIA lacks specific requirements for control, such as access policies or rate limiting. As for the latter, Article 28(2) might even disincentivize efforts to prevent alterations of

the model by exempting the original provider of an AI system from their legal responsibilities when someone along the AI value chain makes a substantial modification to that particular AI system.

With all, the current formulation might be insufficient to rapidly and correctly mitigate harms caused by AI. Through Article 65, the AIA places great responsibility on market surveillance authorities, but their work could be hindered if the provider that first put the AI model into service has not implemented adequate safeguards and monitoring mechanisms. To address this problem, Brakel and Uuk [53] have proposed know-your-customer checks to capture downstream deployers' intended uses and planned modifications to the systems, as well as to urge the original provider to take appropriate action to mitigate identified risks along the value chain. Failing prevention measures of this kind, reversing deployment, if necessary, could become especially hard. Table 2 presents a scrutiny of the AIA based on the discussed policies and the articles that relate to them.

Table 2. Analysis of the requirements set forth in the AIA

Policy	Articles	Diagnosis
Reporting mechanisms	Art. 28b(2e) Art. 28b(2g)	Explicit requirements are lacking but the basis for them exists
Assessment methods	Art. 28b(2a) Art. 28b(2c)	Requirements are appropriate but insufficient and undefined
Post-deployment control	Art. 28b(2e) Art. 65	Important requirements are lacking

Finally, the governance structure proposed in the AIA is arguably the most ambitious to date. It includes several layers, including notified bodies at the micro level, notifying and market surveillance authorities at the national level, and the AI Office at the European level. These institutions will have to surmount serious challenges, including coordination between them and the enforcement of the AIA overseas, where most frontier models are developed. Crucially, they will also have to engage in an ongoing revision of the text and its implementation. The following section covers this issue.

5 Toward a future-proof regulation

Building on the insights from the previous sections, the formulation of the AIA represents a crucial step in the EU's efforts to regulate frontier AI. However, it becomes evident that the journey towards crafting future-proof regulation is challenging due to the rapidly evolving nature of AI. In this section, we discuss whether the AIA is prepared to keep up with that pace through the inclusion of adjustable guidelines and definitions.

The process toward the AIA was officially kick-started with the publication of the *White Paper On Artificial Intelligence* in February 2020 [54], while the Commission presented the first proposal in April 2021 [22]. Since then, the regulation has been reconsidered to incorporate lessons from recent progress in AI. This has translated not only into a reformulation of the AIA's provisions but also of its scope. The inclusion of foundation models is a clear example of this.

Such a constatation unveils an uncomfortable truth: as much as the 2020 proposal is deemed insufficient to address the challenges posed by frontier models, the final AIA is also likely to become outdated in the face of future advancements. This is not new in the elaboration of regulations, which face the challenge of adapting to increasingly rapid technological progress [55], but it becomes especially evident in the frenetic field of AI.

In this context, the AIA could give place to yet another example of flexible technology governance, i.e., a constantly-evolving set of adaptable mechanisms [56]. The European Union could benefit from treating all solutions as incomplete and corrigible, producing an ongoing readjustment of ends and means in a recursive process known as experimentalist governance [57]. The EU is already committed to stress-testing its own policies and making them more resilient to a wider range of low-probability, high-impact events [58].

In fact, the AIA itself is already subject to significant change. First, the interpretation of the definitions, thus the scope of the Regulation might evolve in time. This is exemplified by Article 56b(r), which sets that the AI Office shall "issue and periodically update guidelines on the thresholds that qualify training a foundation model as a large training run".

As for the legal provisions, the text sets forth that the Commission shall "develop, in consultation with the AI Office, guidelines on the practical implementation of [the] Regulation", and "update already adopted guidelines

when deemed necessary" (Article 82b). By virtue of the delegated acts conferred in Article 73, the Commission would also be empowered to update Annex III on high-risk areas and use-cases (Article 7(1)), Annex IV on the technical documentation (Article 11(3)), Annex V on the EU declaration of conformity (Article 48(5)), and Annex VI and VII on conformity assessments (Article 43(5)). Besides, Article 84 stipulates that the Commission shall assess the need for amendment of the list of prohibited AI practices (Article 5) and the list of AI systems requiring additional transparency measures (Article 52). It also states that the Commission "shall, if necessary, submit appropriate proposals to amend [the] Regulation [...] in the light of the state of progress in the information society".

Finally, the establishment of regulatory sandboxes could be an important example of regulatory learning. Following Article 53(5b), "establishing authorities shall submit [...] annual reports [...] [that] provide information on the progress and results of the implementation of those sandboxes, [...] and, where relevant, on the application and possible revision of [the] Regulation".

With all, EU lawmakers seem to be aware of the need to elaborate an adjustable legislation. Notwithstanding, the ability of the AIA to adapt to future technological progress will strongly depend on the performance of European institutions. As concluded by Ibáñez [59] after the analysis of the EU Regulatory Framework for electronic communications, "the success of future-proof regulation does not depend—at least, not primarily—on the *ex-ante* design of a regime, but on the ability of authorities and legislatures to credibly commit, over time, to the same design". As such, "any realistic attempt to design a lasting future-proof regime will have to incorporate mechanisms that make it costly for stakeholders [...] to depart from the approach enshrined in it". If that is the case, the AIA will only be the first step in a long path that will require the timely contribution of the wide range of actors that form the European Union. Further research is then needed to explore how this adaptability could be possible in practice.

6 Conclusion

The AIA represents a pioneering effort in the regulation of frontier models. The law aligns with many of the policy needs identified in recent academic literature, including the establishment of a risk management system or the execution of assessment methods. However, the AIA's effectiveness in addressing the risks associated with frontier models is contingent on the operationalization of its high-level principles and the further standardization of the procedures that it introduces.

Besides, the obligations for providers of foundation models could appear insufficient. First, the AIA does not require the explicit notification of large training runs, even if it does establish the basis for a mechanism like this to exist through the technical documentation and the public database of AI systems. Second, the AIA proposes appropriate practices to assess the models, including evaluations, testing, and an exhaustive identification of risks. Nevertheless, these are not presented clearly as requirements, in contrast with the mandatory conformity assessments that high-risk systems must go through. Third, the AIA presents virtually no obligation to ensure post-deployment control of foundational models. On the contrary, the mitigation of potential damage depends on the performance of market surveillance authorities, whose work can be frustrated if the provider has not previously established the necessary safeguards. Considering those three findings, the AIA would benefit from greater concreteness in the formulation of some obligations and potentially from the inclusion of additional provisions.

Lastly, the AIA marks a substantial advancement in the governance of AI. However, as the technological landscape continues to transform and progress, the regulatory framework will have to adapt accordingly, addressing its own limitations and incorporating novel needs as they emerge. The AIA's success will ultimately depend on the ability of regulators, AI developers, and society at large to engage in a continuous dialogue about the ethical, legal, and societal implications of AI.

Acknowledgements

We would like to thank Risto Uuk and José Villalobos for comments on an earlier draft of this paper. All remaining errors are our own.

References

[1] E. Erdil and T. Besiroglu, 'Algorithmic progress in computer vision', 2022, doi: 10.48550/ARXIV.2212.05153.

[2] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, 'Compute Trends Across Three Eras of Machine Learning', 2022, doi: 10.48550/ARXIV.2202.05924.

[3] M. Anderljung *et al.*, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety', 2023, doi: 10.48550/ARXIV.2307.03718.

[4] T. Shevlane *et al.*, 'Model evaluation for extreme risks', 2023, doi: 10.48550/ARXIV.2305.15324.

[5] OpenAI, 'Frontier Model Forum'. [Online]. Available: https://openai.com/blog/frontier-model-forum

[6] J. Hatzius, J. Briggs, D. Kodnani, and G. Pierdomenico, 'The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani)', 2023. [Online]. Available: https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf

[7] R. Vinuesa *et al.*, 'The role of artificial intelligence in achieving the Sustainable Development Goals', *Nat. Commun.*, vol. 11, no. 1, p. 233, Jan. 2020, doi: 10.1038/s41467-019-14108-y.

[8] J. Jumper *et al.*, 'Highly accurate protein structure prediction with AlphaFold', *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[9] J. Degrave *et al.*, 'Magnetic control of tokamak plasmas through deep reinforcement learning', *Nature*, vol. 602, no. 7897, pp. 414–419, Feb. 2022, doi: 10.1038/s41586-021-04301-9.

[10] G. Liu *et al.*, 'Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii', *Nat. Chem. Biol.*, May 2023, doi: 10.1038/s41589-023-01349-8.

[11] P. Maham and S. Küspert, 'Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks', Stiftung Neue Vertantwortung, 2023.

[12] M. Brundage *et al.*, 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', 2018, doi: 10.48550/ARXIV.1802.07228.

[13] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, 'The Emerging Threat of Ai-driven Cyber Attacks: A Review', *Appl. Artif. Intell.*, vol. 36, no. 1, p. 2037254, Dec. 2022, doi: 10.1080/08839514.2022.2037254.

[14] J. Hazell, 'Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns', 2023, doi: 10.48550/ARXIV.2305.06972.

[15] A. Dezfouli, R. Nock, and P. Dayan, 'Adversarial vulnerabilities of human decision-making', *Proc. Natl. Acad. Sci.*, vol. 117, no. 46, pp. 29221–29228, Nov. 2020, doi: 10.1073/pnas.2016921117.

[16] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks, 'AI Deception: A Survey of Examples, Risks, and Potential Solutions', 2023, doi: 10.48550/ARXIV.2308.14752.

[17] J. B. Sandbrink, 'Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools', 2023, doi: 10.48550/ARXIV.2306.13952.

[18] F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins, 'Dual use of artificial-intelligence-powered drug discovery', *Nat. Mach. Intell.*, vol. 4, no. 3, pp. 189–191, Mar. 2022, doi: 10.1038/s42256-022-00465-9.

[19] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, 'Concrete Problems in AI Safety', 2016, doi: 10.48550/ARXIV.1606.06565.

[20] J. Carlsmith, 'Is Power-Seeking AI an Existential Risk?', 2022, doi: 10.48550/ARXIV.2206.13353.

[21] R. Ngo, L. Chan, and S. Mindermann, 'The alignment problem from a deep learning perspective', 2022, doi: 10.48550/ARXIV.2209.00626.

[22] European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying fown harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. 2021.

[23] Permanent Representatives Committee, *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 - General approach*. 2022.

[24] Committee on the Internal Market and Consumer Protection and Committee on Civil Liberties, Justice and Home Affairs, *Draft Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*. 2023.

[25] C. Siegmann and M. Anderljung, 'The Brussels Effect and Artificial Intelligence', Centre for the Governance of AI, 2022.

[26] D. Owen, 'Extrapolating performance in language modeling benchmarks', 2023.

[27] J. Whittlestone and J. Clark, 'Why and How Governments Should Monitor AI Development', 2021, doi: 10.48550/ARXIV.2108.12427.

[28] Y. Shavit, 'What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring', 2023, doi: 10.48550/ARXIV.2303.11341.

[29] M. Brundage *et al.*, 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims', 2020, doi: 10.48550/ARXIV.2004.07213.

[30] G. Falco *et al.*, 'Governing AI safety through independent audits', *Nat. Mach. Intell.*, vol. 3, no. 7, pp. 566–571, Jul. 2021, doi: 10.1038/s42256-021-00370-7.

[31] J. Schuett, 'AGI labs need an internal audit function', 2023, doi: 10.48550/ARXIV.2305.17038.

[32] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, 'Auditing large language models: a three-layered approach', 2023, doi: 10.48550/ARXIV.2302.08500.

[33] M. Kinniment *et al.*, 'Evaluating Language-Model Agents on Realistic Autonomous Tasks', 2023.

[34] D. Ganguli *et al.*, 'Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned', 2022, doi: 10.48550/ARXIV.2209.07858.

[35] OpenAI, 'GPT-4 Technical Report', 2023, doi: 10.48550/ARXIV.2303.08774.

[36] L. Koessler and J. Schuett, 'Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries', 2023, doi: 10.48550/ARXIV.2307.08823.

[37] A. M. Barrett, D. Hendrycks, J. Newman, and B. Nonnecke, 'Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks', 2022, doi: 10.48550/ARXIV.2206.08966.

[38] M. Anderljung and J. Hazell, 'Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?', 2023, doi: 10.48550/ARXIV.2303.09377.

[39] T. Shevlane, 'Structured access: an emerging paradigm for safe AI deployment', 2022, doi: 10.48550/ARXIV.2201.05159.

[40] J. Wei *et al.*, 'Emergent Abilities of Large Language Models', 2022, doi: 10.48550/ARXIV.2206.07682.

[41] J. Wei *et al.*, 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models', 2022, doi: 10.48550/ARXIV.2201.11903.

[42] Y. Liu *et al.*, 'Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study', 2023, doi: 10.48550/ARXIV.2305.13860.

[43] I. Solaiman and C. Dennison, 'Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets', 2021, doi: 10.48550/ARXIV.2106.10328.

[44] L. Ouyang *et al.*, 'Training language models to follow instructions with human feedback', 2022, doi: 10.48550/ARXIV.2203.02155.

[45] J. Schuett *et al.*, 'Towards best practices in AGI safety and governance: A survey of expert opinion', 2023, doi: 10.48550/ARXIV.2305.07153.

[46] K. Zenner, 'A law for foundation models: the EU AI Act can improve regulation for fairer competition', *OECD*, Jul. 20, 2023. [Online]. Available: https://oecd.ai/en/wonk/foundation-models-eu-ai-act-fairer-competition

[47] L. Bertuzzi, 'Spanish presidency pitches obligations for foundation models in EU's AI law', *EURACTIV*, Nov. 07, 2023. [Online]. Available: https://www.euractiv.com/section/artificial-intelligence/news/spanish-presidency-pitches-obligations-for-foundation-models-in-eus-ai-law/

[48] P. Cihon, J. Schuett, and S. D. Baum, 'Corporate Governance of Artificial Intelligence in the Public Interest', *Information*, vol. 12, no. 7, Art. no. 7, Jul. 2021, doi: 10.3390/info12070275.

[49] A. Askell, M. Brundage, and G. Hadfield, 'The Role of Cooperation in Responsible AI Development', 2019, doi: 10.48550/ARXIV.1907.04534.

[50] J. Mökander, M. Axente, F. Casolari, and L. Floridi, 'Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation', *Minds Mach.*, vol. 32, no. 2, pp. 241–268, Jun. 2022, doi: 10.1007/s11023-021-09577-4.

[51] I. Solaiman, 'The Gradient of Generative AI Release: Methods and Considerations', 2023, doi: 10.48550/ARXIV.2302.04844.

[52] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, 'Universal and Transferable Adversarial Attacks on Aligned Language Models', 2023, doi: 10.48550/ARXIV.2307.15043.

[53] M. Brakel and R. Uuk, 'FLI Position Paper: AI Act Trilogue', Future of Life Institute, 2023.

[54] European Commission, 'White Paper on Artificial Intelligence: a European approach to excellence and trust', European Commission, Brussels, White Paper COM(2020) 65 final, Feb. 2020. [Online]. Available: https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[55] D. Collingridge, *The social control of technology*. New York: St. Martin's Press, 1980.

[56] R. Hagemann, J. Huddleston, and A. Thierer, 'Soft Law for Hard Problems: The Governance of Emerging Technologies in an Uncertain Future', *SSRN Electron. J.*, Feb. 2018.

[57] C. F. Sabel and J. Zeitlin, *Experimentalist Governance*. Oxford University Press, 2012. doi: 10.1093/oxfordhb/9780199560530.013.0012.

[58] M. Fernandes and A. Heflich, '"Future proofing" EU policies. The why, what and how to stress testing', European Parliament, Briefing, 2021.

[59] P. Ibáñez Colomo, 'Future-Proof Regulation against the Test of Time: The Evolution of European Telecommunications Regulation', *Oxf. J. Leg. Stud.*, vol. 42, no. 4, pp. 1170–1194, Dec. 2022, doi: 10.1093/ojls/gqac016.