RCG POSITION PAPER: AI ACT TRILOGUE

RIESGOS CATASTRÓFICOS GLOBALES

October 2023

Guillem Bas
gbasg@riesgoscatastroficosglobales.com

This position paper advances six recommendations for policymakers involved in the negotiations toward the EU AI Act. It includes four proposals to regulate frontier models, a suggestion for the nature and tasks of the future institution established by the Act, and a section on how to address open-source models. The first page compiles, for each proposal, the specific paragraph that could be included in the regulation. The rest of the paper justifies and develops on the six recommendations.

—

RCG is an international science-policy organisation working on the formulation of governance proposals that reduce risks related to the development and deployment of advanced artificial intelligence. To achieve our mission, we connect policymakers with experts and produce reports with evidence-based recommendations.

Recommendations

1. Frontier models

a. Compute thresholds

'Frontier model' means a highly capable foundation model, resulting from a training process above a threshold defined by benchmarking authorities, that could possess dangerous capabilities sufficient to pose severe risks to public safety [1]

b. Third-party model evaluations and testing

Design and develop the frontier model to achieve appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity throughout its lifecycle, as examined through model evaluation conducted by independent experts, documented analysis, and extensive testing during design, development, testing, and deployment

c. Risk management system

Demonstrate, through appropriate design, testing and analysis throughout the entire lifecycle of the frontier model and with the involvement of independent experts, that known and reasonably foreseeable risks to health, safety, fundamental rights, the environment, and democracy and the rule of law are eliminated, reduced, or mitigated to an overall residual risk reasonably judged to be acceptable, while facilitating the documentation of remaining non-mitigable risks after development

d. Deployment safeguards

When providing the frontier model as a service such as through an API, retain control over the access and use of the model, keep and monitor the logs automatically generated by the provided interface to identify any instances of serious malfunction, incidents, or misuse, and prepare prevention and contingency plans that facilitate the execution of corrective actions in the event of such cases.

2. Governance

Establish an AI Office as an independent body with legal personality that shall, among others, issue thresholds to define frontier models, act as a hub for evaluation efforts, and conduct assessments of large-scale risks.

3. Open source

A provider of a frontier model shall ensure compliance with the Regulation regardless of whether it is provided under free and open source licenses.

[1] For further consideration on how to define frontier models, see Anderljung et al. (2023), Appendix A.

Compute thresholds

The AI Act should introduce a distinct subset of foundation models called *frontier models*, initially defined by compute-based thresholds, that would be subject to particularly stringent requirements.

By introducing a tiered approach to foundation models (FMs) or general-purpose AI systems (GPAIS), the AI Act would relieve concerns about some obligations being too burdensome for smaller providers of those models or systems. In fact, narrowing down the scope to a handful of providers could justify even more stringent obligations.

Computational resources (compute) used during the training process of a model constitute a promising variable to define frontier models, since there exists a strong correlation between training compute and the performance of the resulting model (Owen, 2023). As such, compute could be a relatively effective indicator of the emergence of strong capabilities that, due to their dual nature, deserve careful oversight. Besides, compute gathers several advantages as a governance lever, as it is easily measurable (Sevilla et al., 2023) and monitorable (Shavit, 2023).

Some elements in the AI Act could serve as a basis to establish compute thresholds for FMs or GPAIS. For instance, the draft proposed by the European Parliament (EP) requires providers of FMs to provide, as part of the technical documentation, 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". Such information could be included in the registration of the model in the EU database to increase regulatory visibility and facilitate a compute-based classification of models.

An initial threshold could be 1e24 floating-point operations (FLOP), above which there are only seven current models (Epoch, 2022). Arguably, these models might pose sufficiently significant risks to deserve special attention, while a lower threshold might be harder to enforce and include less risk-relevant models. However, these thresholds should be periodically updated according to a downward force—future algorithmic progress that would make certain capabilities accessible with less compute (Erdil & Besiroglu, 2022)—and an upward force—our increasing ability to understand existing models, their risk profile, and how to manage them, all of which would deprioritize governance of models as they get farther from the frontier. The EP's text tasks the AI Office to "issue and periodically update guidelines on the thresholds that qualify training a foundation model as a large training run", which would be especially relevant if those thresholds have regulatory implications.

Nevertheless, regulators should consider the limitations of such proposal, including that (i) compute costs halve roughly every 30 months (Hobbhahn & Besiroglu, 2022); (ii) many individual tasks are harder to relate to compute scale (Owen, 2023) and some have been shown to scale inversely with training compute (McKenzie et al., 2023); and (iii) some narrow systems with lower compute requirements, such as models for drug discovery (Urbina et al., 2022) or AI-powered malware (Sims, 2023), also carry significant risks. Accordingly, compute thresholds should be complemented with other benchmarks and, in any case, only be part of a more nuanced governance framework.

Third-party model evaluations and testing

The AI Act should explicitly require developers of frontier models to carry out model evaluations and extensive testing led by independent auditors.

To ensure public safety, these evaluations should focus on dangerous capabilities—to what extent a model is capable of causing extreme harm—and alignment—to what extent a model has the propensity to cause extreme harm (Shevlane et al., 2023). Important examples of dangerous capabilities that evaluations should examine include autonomous replication and adaptation (Kinniment et al., 2023) and deception of humans (OpenAI, 2023). In any case, evaluations should be standardized so that they are consistently applicable across models and rely as little as possible on the auditor's or auditee's discretion (Anderljung et al., 2023).

Evaluations would help inform if the model can be deployed and how (Shevlane et al., 2023). In the EU, the output of such evaluations should be crucial to assess if the model is safe to be commercialized and compliant with the AI Act. However, it is important to ensure that the evaluation not only assesses the capabilities of the raw model, but also anticipates all foreseeable scenarios that could help expand the capabilities of the model after deployment, such as fine-tuning (Yoosuf & Yang, 2019) or its integration into agential programs (Xi et al., 2023). Though harder to enforce, EU institutions should also promote evaluations during training as an example of best practice, as these evaluations would help identify dangerous capabilities that emerge unexpectedly after a certain threshold (Wei et al., 2022). These practices could in turn enable decisions such as discontinuing training or deleting model weights to prevent the proliferation of dangerous capabilities through leaks or theft.

As a complementary effort, models should be subject to red teaming, i.e., a structured effort to find vulnerabilities and flaws in the model, performed by a 'red team' that adopts an attacker's mindset and methods (Brundage et al., 2020). Red teaming exercises have been used by frontier labs to anticipate and correct undesired output such as information to develop weapons or conduct cyberattacks, misinformation, discrimination, hate speech, and incitement, among others (Ganguli et al., 2022; OpenAI, 2023). However, current models are still vulnerable to 'jailbreaks', i.e., prompts that circumvent model constraints and elicit unintended behaviour (Liu et al., 2023; Zou et al., 2023). For that reason, stronger testing is needed to ensure the reliability, robustness, and cybersecurity of future models.

Model evaluations and red teaming exercises should be carried out by third-party actors to avoid conflicts of interest (Brundage et al., 2020). Model evaluations and red teaming exercises are among the most favoured practices for experts and practitioners (Schuett et al., 2023).

Risk management system

The AI Act should require providers of foundation models to implement a risk management system composed of (i) an exhaustive risk assessment and (ii) standardized protocols for the elimination or mitigation of risks.

Exhaustive risk assessment

Assessment of extreme risks would mainly consist of model evaluations and red teaming exercises, which are covered in the previous section. However, there are other techniques of risk identification (e.g., scenario analysis and fishbone method), analysis (e.g., causal mapping and Delphi technique), and evaluation (e.g., checklists and matrices) that developers of frontier models could borrow from other safety-critical industries (Koessler & Schuett, 2023).

Risk assessment processes should be carried out during the entire lifecycle of the model, including design, development, testing, and deployment. Decisions made at each stage should be informed by these assessments.

Later standards and specifications of the AI Act should clarify the level of diligence required to assess risk, i.e., the amount of effort a provider must devote to the process to ensure that all relevant risks are identified and foreseen (Schuett, 2023). However, risk assessment processes should be designed so that severe and catastrophic risks and longer-term impacts enter in their scope (Barrett et al., 2022).

Protocols for the elimination or reduction of risks

During development, risk assessments might uncover the need to adjust the model to circumvent the identified risks. Currently, adjustments are commonly done through techniques such as fine-tuning (Solaiman & Dennison, 2021) or reinforcement learning from human feedback (Ouyang et al., 2022).

Once the model is trained, deployment rules should be proportionate to the model's risk profile, i.e., guardrails should be stronger the more severe and uncertain risks are (Anderljung et al., 2023). An emerging paradigm in that regard is responsible scaling policy, which aims to keep protective measures ahead of whatever dangerous capabilities AI models have. In other words, it is a commitment to pause deployment and/or development of a model if specific dangerous capabilities emerge, until protective measures are good enough to handle them safely (ARC Evals, 2023). Some authors even argue that, in the event of such a scenario, other labs should also pause their activities with models of similar scale (Alaga & Schuett, 2023).

Deployment safeguards

The AI Act should require providers of frontier models to keep control over their models along the entire value chain and prepare plans to prevent or quickly respond to incidents involving deployed models.

First, frontier models should be deployed gradually to prevent failures related to a sudden shift in the environment (Amodei et al., 2016) or facilitating early access to the model to a wide group of users, increasing the potential for harmful use cases (Solaiman, 2023). Some proposals for a gradual deployment include structured access, a controlled arm's length interaction between the model and its user (Shevlane, 2022), and the gradient of AI release, which lays down six levels of progressive access to navigate the tradeoffs between concentrating power and mitigating risks (Solaiman, 2023).

Second, providers should leverage that control over their models by monitoring downstream uses through know-your-customer checks (Brakel & Uuk, 2023), always within personal data protection limits (Bluemke et al., 2023). Some frontier labs already retain API inputs and outputs between 30 and 90 days to identify potential misuse (Anthropic, n.d.; OpenAI, n.d.). In the context of the AI Act, an obligation to monitor API usage could be inspired by Article 20, which requires providers of high-risk systems to keep the logs automatically generated by their systems.

Finally, it is essential that providers of frontier models adopt security measures to prevent or terminate downstream misuses (Anderljung & Hazell, 2023). In this context, possible deployment corrections include user-based restrictions, access frequency limits, capability or feature restrictions, use case restrictions, and shutdown (O'Brien et al., 2023). Similarly, providers should build out the tooling, policies, procedures, and roles necessary for an effective incident response, which could learn from other safety-critical industries such as the field of cybersecurity. These ideas go in line with Article 21, which requires providers of high-risk systems to take the necessary corrective actions to reconduct a deployed system that does not comply with the regulation.

None of the proposed versions of the AI Act have given sufficient importance to the obligations of providers of FMs or GPAIS after placing them in the market or putting them into service. However, if providers do not comply with a minimum set of obligations in that regard, such as establishing a post-market monitoring system or reporting serious incidents, market surveillance authorities might have a hard time managing incidents derived from these cutting-edge technologies, as established in Title VIII.

In fact, some parts of the AI Act might disincentivize original providers from assuming legal responsibility for their models along the value chain, as they would cease to be providers in cases in which downstream deployers market the model under their name or substantially modify the model. A joint and several liability regime would be more appropriate to incentivize shared responsibility and ensure that the original provider remains available to intervene if it becomes necessary to do so. This would also incentivize original providers to support downstream providers in the re-evaluation of the model after a substantial modification that could change the model's risk profile.

Governance

The AI Act should establish an AI Office as an independent body with legal personality that shall, among others, ensure an effective governance of frontier models.

An AI Office, understood as an independent body with legal personality as opposed to a mere coordination mechanism like the AI Board, stands out as a particularly effective, efficient, and coherent institution model to enforce the AI Act internally and abroad (Moës et al., 2023). A well-resourced agency can become the main point of contact for EU stakeholders and a relevant counterpart to key jurisdictions and providers on the international scene.

As for the functions that the proposed AI Office and AI Board would assume, there are several coincident tasks, such as pooling technical expertise, coordinating and assisting Member states, or supporting the Commission in relation to the implementation of the AI Act. However, the AI Office proposed by the EP includes some additional tasks that would be relevant for the governance of frontier models, including:

- Article 56b(o): "provide monitoring of foundation models and to organise a regular dialogue with the developers of foundation models with regard to their compliance as well as AI systems that make use of such AI models."
- Article 56b(q): "provide particular oversight and monitoring and institutionalize regular dialogue with the providers of foundation models about the compliance of foundation models [...] and about industry best practices for self-governance"

To enhance this oversight, the AI Act could mirror the Digital Services Act and require providers of frontier models to designate a compliance function, i.e., mandating them to establish specific roles to ensure compliance with the regulation (Moës et al., 2023).

In line with the previous proposals, the AI Office should also:

- Issue and periodically update guidelines on the thresholds defining frontier models, as already proposed in the EP's version (Article 56b(r)).
- Act as a hub for nascent evaluation efforts by coordinating between private companies, regulators, and sector-specific experts (Apollo Research, 2023). Furthermore, model evaluators should report their outcomes to the AI Office and the Commission, in the same way as notified bodies and market surveillance authorities.
- Conduct assessments of large-scale risks posed by frontier models, potentially as part of its obligation to "issue an annual report on the state of play in the development, proliferation, and use of foundation models" (Art. 56b(r)). This would also align with the Commission's call to conduct more comprehensive assessments of risks to economic security (European Commission, 2023).

Besides enforcement, the AI Office should assist other institutions in their efforts to refine the AI Act after its approval. More specifically, the new body could contribute to standardization by identifying priority areas and support the Commision in the elaboration of implementing and delegated acts (Pouget & Laux, 2023). To be 'future-proof', the AI Act will need to be continuously adapted to the rapidly changing field of AI, so its effectiveness will largely depend on the agility of the involved authorities to keep it updated.

Open source

The AI Act should not exempt open-access[2] frontier models from the aforementioned obligations, keeping providers partially liable for potential harms caused by these open-sourced frontier models.

Both the Council's and the EP's proposals state that general-purpose AI systems or foundation models shall be compliant with the regulation regardless of whether they are provided under free and open source licenses. These clarifications are important to prevent providers from exploiting an exemption and therefore bypass the regulation.

Arguably, requirements such as model evaluations and risk assessments are especially important for open-access frontier models, as providers lose control over their functioning and therefore must be especially confident that their models will not cause unacceptable harms in any case. In other words, it is important to ensure that no unresolved safety issues and no possible misuse can cause serious incidents, as the provider's capacity to intervene in one of those incidents is much smaller when the model is open-access (Seger et al., 2023).

With all, providers of open-access frontier models should be incentivized to comply with the established requirements before publishing the model because, failing that, the model could not be legally put into service or in the market by other distributors. Furthermore, despite not having a formal relationship, operators integrating an open-access model into their systems should be able to report incidents to the upstream developer, who would have the obligation to support the downstream provider's efforts to bring that AI system into compliance or withdraw it.

As for non-frontier, open-source foundation models, these could be exempted from the aforementioned requirements to leverage their benefits, such as enabling external oversight, advancing safety research, or distributing influence and benefits (Seger et al., 2023). For example, open access to models can facilitate the detection of flaws in training datasets (Piktus et al., 2023) or improve the assessment of a model's performance (Fourrier et al., 2023), while many advances in fields like mechanistic interpretability have been possible thanks to independent research based on open-source models (Conmy et al., 2023; Wang et al., 2022). Besides, open-source permits the existence of large collaborative projects that allow researchers to leverage the technology without depending on well-resourced corporations (Workshop et al., 2022).

[2] In this context, an 'open-access' model means one that is openly released by its provider, which involves "making model architecture and weights freely and publicly accessible for anyone to modify, study, build on, and use" (Seger et al., 2023). In this case, however, some components such as the training dataset might not be downloadable, and the overall accessibility of the model may be limited by high computing requirements and commercialization restrictions (Hull, 2023; Solaiman, 2023). Note that this is different from the model being 'open-source', i.e., released under a licence in which "the copyright holder grants users the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose" (Corbly, 2014).

References

Alaga, J., & Schuett, J. (2023). *Coordinated pausing: An evaluation-based coordination scheme for frontier AI developers*. https://doi.org/10.48550/ARXIV.2310.00374

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. https://doi.org/10.48550/ARXIV.1606.06565

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). *Frontier AI Regulation: Managing Emerging Risks to Public Safety*. https://doi.org/10.48550/ARXIV.2307.03718

Anderljung, M., & Hazell, J. (2023). *Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?* https://doi.org/10.48550/ARXIV.2303.09377

Anthropic. (n.d.). *How long do you store personal data?* https://support.anthropic.com/en/articles/7996866-how-long-do-you-store-personal-data

Apollo Research. (2023, October 11). *Recommendations for the next stages of the Frontier AI Taskforce*. https://www.apolloresearch.ai/blog/recommendations-for-the-frontier-ai-taskforce

ARC Evals. (2023, September 26). *Responsible Scaling Policies (RSPs)*. https://evals.alignment.org/blog/2023-09-26-rsp/

Barrett, A. M., Hendrycks, D., Newman, J., & Nonnecke, B. (2022). *Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks*. https://doi.org/10.48550/ARXIV.2206.08966

Bluemke, E., Collins, T., Garfinkel, B., & Trask, A. (2023). *Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases*. https://doi.org/10.48550/ARXIV.2303.08956

Brakel, M., & Uuk, R. (2023). *FLI Position Paper: AI Act Trilogue*. Future of Life Institute.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. https://doi.org/10.48550/ARXIV.2004.07213

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). *Towards Automated Circuit Discovery for Mechanistic Interpretability*. https://doi.org/10.48550/ARXIV.2304.14997

Corbly, J. E. (2014). The Free Software Alternative: Freeware, Open Source Software, and Libraries. *Information Technology and Libraries*, *33*(3), 65. https://doi.org/10.6017/ital.v33i3.5105

Epoch. (2022). *Parameter, Compute and Data Trends in Machine Learning*. https://epochai.org/mlinputs/visualization

Erdil, E., & Besiroglu, T. (2022). *Algorithmic progress in computer vision*. https://doi.org/10.48550/ARXIV.2212.05153

European Commission. (2023, June 20). *An EU approach to enhance economic security* *. https://ec.europa.eu/commission/presscorner/detail/en/IP_23_3358

Fourrier, C., Habib, N., Launay, J., & Wolf, T. (2023, June 23). What's going on with the Open LLM Leaderboard? *Hugging Face*. https://huggingface.co/blog/evaluating-mmlu-leaderboard

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., … Clark, J. (2022). *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*. https://doi.org/10.48550/ARXIV.2209.07858

Hobbhahn, M., & Besiroglu, T. (2022). Trends in GPU price-performance. *Epoch*. https://epochai.org/blog/trends-in-gpu-price-performance

Hull, C. (2023, July 19). Is Llama 2 open source? No – and perhaps we need a new definition of open…. *OpenSource Connections*.

https://opensourceconnections.com/blog/2023/07/19/is-llama-2-open-source-no-and-perhaps-we-need-a-new-definition-of-open/
Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., Ho, A., Barnes, E., & Christiano, P. (2023). *Evaluating Language-Model Agents on Realistic Autonomous Tasks*.
Koessler, L., & Schuett, J. (2023). *Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries*. https://doi.org/10.48550/ARXIV.2307.08823
Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., & Liu, Y. (2023). *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study*. https://doi.org/10.48550/ARXIV.2305.13860
McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., Gritsevskiy, A., Wurgaft, D., Kauffman, D., Recchia, G., Liu, J., Cavanagh, J., Weiss, M., Huang, S., Droid, T. F., … Perez, E. (2023). *Inverse Scaling: When Bigger Isn't Better*. https://doi.org/10.48550/ARXIV.2306.09479
Moës, N., Reddel, F., & Curtis, S. (2023). *Giving Agency to the EU AI Act Comparing Options for Enforcement*. The Future Society.
O'Brien, J., Ee, S., & Williams, Z. (2023). *Deployment corrections: An incident response framework for frontier AI models*. Institute for AI Policy and Strategy.
OpenAI. (n.d.). *Enterprise privacy at OpenAI*. https://openai.com/enterprise-privacy
OpenAI. (2023). *GPT-4 Technical Report*. https://doi.org/10.48550/ARXIV.2303.08774
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. https://doi.org/10.48550/ARXIV.2203.02155
Owen, D. (2023). *Extrapolating performance in language modeling benchmarks*.
Piktus, A., Akiki, C., Villegas, P., Laurençon, H., Dupont, G., Luccioni, A. S., Jernite, Y., & Rogers, A. (2023). *The ROOTS Search Tool: Data Transparency for LLMs*. https://doi.org/10.48550/ARXIV.2302.14035
Pouget, H., & Laux, J. (2023, October 3). A Letter to the EU's Future AI Office. *Carnegie Endowment for International Peace*. https://carnegieendowment.org/2023/10/03/letter-to-eu-s-future-ai-office-pub-90683
Schuett, J. (2023). Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, 1–19. https://doi.org/10.1017/err.2023.1
Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., & Garfinkel, B. (2023). *Towards best practices in AGI safety and governance: A survey of expert opinion*. https://doi.org/10.48550/ARXIV.2305.07153
Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., Ó hÉigeartaigh, S., Korinek, A., Anderljung, M., Bucknall, B., Chan, A., Stafford, E., Koessler, L., Ovadya, A., Garfinkel, B., Bluemke, E., Aird, M., … Gupta, A. (2023). *Open-Sourcing Highly Capable Foundation Models. An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*.
Sevilla, J., Ho, A., & Besiroglu, T. (2023). Please Report Your Compute. *Communications of the ACM*, *66*(5), 30–32. https://doi.org/10.1145/3563035
Shavit, Y. (2023). *What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring*. https://doi.org/10.48550/ARXIV.2303.11341
Shevlane, T. (2022). *Structured access: An emerging paradigm for safe AI deployment*. https://doi.org/10.48550/ARXIV.2201.05159
Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023). *Model evaluation for extreme risks*. https://doi.org/10.48550/ARXIV.2305.15324
Sims, J. (2023, August 2). EyeSpy Proof-of-Concept. *HYAS*. https://www.hyas.com/blog/eyespy-proof-of-concept
Solaiman, I. (2023). *The Gradient of Generative AI Release: Methods and Considerations*.

https://doi.org/10.48550/ARXIV.2302.04844
Solaiman, I., & Dennison, C. (2021). *Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets*. https://doi.org/10.48550/ARXIV.2106.10328
Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189–191. https://doi.org/10.1038/s42256-022-00465-9
Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). *Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small*. https://doi.org/10.48550/ARXIV.2211.00593
Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models*. https://doi.org/10.48550/ARXIV.2206.07682
Workshop, B., :, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., … Wolf, T. (2022). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. https://doi.org/10.48550/ARXIV.2211.05100
Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Liu, Q., Zhou, Y., Wang, W., Jiang, C., Zou, Y., … Gui, T. (2023). *The Rise and Potential of Large Language Model Based Agents: A Survey*. https://doi.org/10.48550/ARXIV.2309.07864
Yoosuf, S., & Yang, Y. (2019). Fine-Grained Propaganda Detection with Fine-Tuned BERT. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 87–91. https://doi.org/10.18653/v1/D19-5011
Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and Transferable Adversarial Attacks on Aligned Language Models*. https://doi.org/10.48550/ARXIV.2307.15043