Foundation Models and the EU AI Act

Rishi Bommasani
Stanford CRFM

Alice Hau
Stanford HAI

Kevin Klyman
Stanford CRFM

Percy Liang
Stanford CRFM

Abstract

The EU AI Act is the world's first comprehensive legal regime for governing artificial intelligence. The Act reflects several years of legislative process in the European Union and, in particular, laws that grapple with the emerging technology of foundation models. We analyze how the Act addresses foundation models by coding the Act's 31 requirements for foundation model developers into a multi-level taxonomy: 87% are disclosures, yet only one requirement requires information be disclosed publicly. Using our coding as a lens, we juxtapose the AI Act with prior legislative proposals. While the proposals and the Act emphasize transparency, the Act lacks the public-facing transparency sought in previous proposals. While time will tell how the EU AI Act shapes global AI development and policymaking, our work helps to set expectations for its handling of foundation models.

1 Introduction

The growing importance of artificial intelligence (AI) catalyzes global policymaking efforts. Foremost of all these efforts is the European Union's AI Act, which constitutes the first set of binding laws to comprehensively address AI. The AI Act continues an established tradition of EU policymaking on the digital technologies, evinced by GDPR (data privacy), the DSA (online platforms), and the DMA (digital markets). The AI took roughly five year from its formulation in 2020 through its negotiation in 2023 and its enactment in 2024. The AI Act will significantly impact not just EU, but also global, AI development and AI policy. In particular, it sets the definitive precedent for how policy will reckon with this fast-moving technology, though only time will reveal the nature of this impact.

In this work, we analyze the EU AI Act with a focus on foundation models and general-purpose AI. *General-purpose AI models* (GPAI models) are defined by the Act as "AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks". GPAI models, which are generally referred to as *foundation models* [Bommasani et al., 2021], have powered recent advances in artificial intelligence. Models like OpenAI's GPT-4, Meta's Llama 3, Google's Gemini 1.5 and Anthropic's Claude 3.5 are highly capable, resource intensive, and widely adopted. These models, and their developers, play an outsized role in defining the AI supply chain and mediating AI's sweeping societal and economic impact. In turn, the Act's treatment of these cutting-edge models reflects a cutting-edge policy approach worthy of deeper inspection.

To understand the AI Act in relation to foundation models, we *code* the obligations imposed upon their providers. Our coding organizes the obligations based on whether they require information disclosures or substantive actions, target all developers or specific subsets, and categorize who receives any disclosed information (e.g. the public vs. the EU AI Office). We count 31 requirements that apply to 4 classes of general-purpose artificial intelligence. Most requirements are information disclosures where developers provide information to either the government or downstream firms (84%). The four substantive requirements that compel a developer to take a particular action hinge on whether a model is designated as posing systemic risk. As of August 2024, we find that 8 models meet the default systemic risk criterion of 10^{25} training FLOPs based on the estimates of Epoch

2nd Workshop on Regulatable ML at NeurIPS 2024.

AI [2024], while only one model (OpenAI's GPT-4) would have qualified at the time of legislative agreement in December 2023.

The foundation model requirements will go into effect on August 2, 2025, with the EU AI Office currently working on designing codes of practice to facilitate provider compliance. In the interim, we look back at the Act's legislative process as an additional lens. In particular, we consider two proposals made during the Act's negotiation in 2023. The first is from the European Parliament: this proposal was adopted by Parliament based on a decisive vote in June 2023 and functioned as its negotiating position in the AI Act negotiations in June to December 2023. The second is from Stanford researchers [Bommasani et al., 2023]: this proposal was put forth as a concrete solution for achieving compromise between the different EU legislators on December 1, 2023, just six days before when the Act was successfully negotiated on December 7.

We code the Parliament and Stanford proposals to facilitate a comparative analysis. We find that each of the three texts (i.e. the two proposals and the AI Act) primarily centers on information disclosure: the Parliament position and the Stanford proposal emphasize public-facing disclosure, whereas the AI Act includes just one public-facing disclosure requirement that is specific to general-purpose AI. The Stanford proposal and the AI Act share a two-tiered approach and consider many similar criteria for setting the threshold, yet the AI Act's default criterion of compute diverges from the Stanford proposal's focus on demonstrated market impact. Critically, while the AI Act includes many of the elements of the Stanford proposal, it lacks provisions for third-party researcher access to models. The non-inclusion is particular interesting given that the EU guaranteed third-party researcher access for online platforms in the Digital Services Act. Through this juxtaposition, we identify key factors (e.g. limited public-facing transparency) that may limit the extent to which the EU AI Act significantly increases AI accountability.

2 Coding the EU AI Act

The legal text for the EU AI Act is an expansive 144 page document with 113 articles and 12 annexes. Overall, the Act takes a risk-based approach to AI governance, classifying AI systems into four categories based on the application area (prohibited, high-risk, limited risk, and minimal risk). Since foundation models are general-purpose technologies [Eloundou et al., 2024] that can be applied in many ways, these models are subject to a separate set of obligations.

Chapter V on General-Purpose AI Models (Articles 51 – 56), along with three supporting annexes (Annex XI – XIII), specifies the obligations for the providers of foundation models. These obligations vary significantly based on whether a model is classified as posing *systemic risk* (Articles 51 – 52). While the obligations for general-purpose AI take effect on August 2, 2025, the EU AI Office is tasked (Article 56) with preparing codes of practice by May 2, 2025 to facilitate compliance.

Coding schema. Given the Act's complexity, we organize the requirements by coding them. To begin, we broke down large requirements into atomic units that we label with a short descriptor. For example, there are several elements that model providers are required to documents such as information on model capabilities, evaluations, and risk mitigation. This yielded 25 requirements.

For each requirement, we tag the *type*, the *recipient*, and the *scope*. For the foundation model requirements, every requirement is typed as either a*substantive* (i.e. the provider is required to implement a specific practice) or a *disclosure* (i.e. the provider is required to disclose information to some other party) requirement. For the subset of disclosure requirements, every requirement is tagged based on whether information must be disclosed to the public, to downstream firms that will integrate the foundation model, or to the government (i.e. the EU AI Office and national governments). For the scope, requirements could apply to (i) all providers, (ii) providers of open models without systemic risk, (iii) providers of non-open models without systemic risk, or (iv) providers of models with systemic risk.

Takeaway 1: Overall focus on transparency. Our coding makes clear that the focus of the EU AI Act requirements for foundation model providers is on increased transparency through information disclosure. 27 (87%) of the requirements are disclosures. These disclosures are primarily directed towards the government or to downstream firms. In sharp contrast, there is only one public-facing disclosure requirement. Consequently, in spite of the emphasis on increased information sharing, the benefits of this disclosure are less certain given that researchers, journalists, and the broader public may not have access to information, in spite of their vital role in the accountability ecosystem.

The sole public-facing disclosure requirement states that providers must "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model". This first-of-its-kind requirement could substantially increase overall transparency into training data, given the documented opacity specifically for model training data [Bommasani et al., 2024]. In turn, the EU AI Office's template for this training data summary is a critical focus for determining what increased insight and accountability will come about due to this requirement. Warso et al. [2024] put forth a concrete and extensive proposal for what should be required.

Takeaway 2: Substantive requirements mostly hinge on systemic risk designation. In contrast to the large number of disclosure requirements, there are just four (13%) substantive requirements. Only one of these requirements, which requires providers to implement a policy to ensure adherence with copyright law, is required for all general-purpose AI models. The remaining three substantive requirements are only required for models designated as posing systemic risk.

The specific substantive obligations pertain to conducting model evaluations, mitigating possible risks, and implementing cybersecurity protections. As with the training data summary, the legislative language of the EU AI Act is significantly underspecified: for example, the requirement for model evaluations states that providers should "perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art". This introduces legal ambiguity on what technical actions would constitute compliance. Therefore, much as with the training data summary requirement, the work of the EU AI Office to clarify these expectations through their forthcoming codes of practice will be formative. The focus on evaluations, mitigations, and cybersecurity is also resonant with policy in other jurisdictions such as the US Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [Executive Order 14110, 2023] and the G7 International Code of Conduct for Organizations Developing Advanced AI Systems [Group of Seven, 2023].

Takeaway 3: Several models may be subject to compute-based systemic risk designation. The default basis for designating a model as posing systemic risk is whether the cumulative training compute eclipses 10^{25} FLOPs. In recognition of several known critiques of compute [Hooker, 2024], Annex XIII identifies alternative metrics the AI Office may consider such as data properties, evaluation results, number of business users, and number of registered end-users [see Bommasani, 2023]. In addition, the AI Act also creates a scientific panel of efforts that is empowered to alert the AI Office of models that may pose systemic risk, whether that is because they eclipse the compute threshold or for other reasons. Given that many of the most substantial requirements under the AI Act are apportioned to models posing systemic risk, the established thresholds and scientific panel's actions will heavily shape compliance burdens.

Today, many of the most prominent foundation models are released without public disclosure of the amount of training compute [Bommasani et al., 2024]. Therefore, to understand the current threshold of 10^{25} FLOPs, we turn to Epoch database on large models [Epoch AI, 2024], which estimates the compute for models and is widely recognized as the most authoritative public resource on the matter [Bengio et al., 2024]. As of August 2024, the database indicates that 8 models (Gemini 1.0 Ultra, Llama 3.1-405B, GPT-4, Mistral Large, Nemotron-4 340B, MegaScale, Inflection-2, Inflection-2.5) from 7 developers (Google, Meta, OpenAI, Mistral, NVIDIA, ByteDance, Inflection) were trained with a (potentially estimated) compute exceeding 10^{25} FLOPs. This is particularly striking given that at the time of legislative agreement on the EU AI Act in December 2023, only one model (GPT-4 from OpenAI) met this criterion, indicating that the pace at which these thresholds are modified will be quite critical.

Takeaway 4: Open foundation models receive a partial exemption. Article 53 states that many of the obligations "do not apply to providers of AI models that are released under a free and open-source licence", but that "this exception shall not apply to general-purpose AI models with systemic risks". Further, the preamble of the Act explicitly justifies why open-source is subject to fewer requirements, articulating that open-source can "contribute to research and innovation in the market and can provide significant growth opportunities for the [European] Union economy". These views align with the demonstrated track record for open-source software [Blind et al., 2021, Hoffmann et al., 2024] as well as work from both scientists [Kapoor et al., 2024, Vipra and Korinek, 2023] and commerce agencies in other jurisdictions [UK Competition and Markets Authority, 2023, US Federal Trade Commission, 2024, US National Telecommunications and Information Administration, 2024].

The specific interpretation of what constitutes a "free and open-source" license is therefore consequential: the Act clarifies this as "when it allows users to run, copy, distribute, study, change and improve software and data, including models under the condition that the original provider of the model is credited, the identical or comparable terms of distribution are respected". In particular, the term open-source has a widely accepted definition for code, maintained by the Open Source Initiative (OSI), but no established counterpart for artificial intelligence or model weights. The OSI is currently working to define open-source artificial intelligence,[1] but we highlight that many foundation models with widely available model weights are released under licenses that do not comply with the OSI standard for open-source software (e.g. BLOOM, Stable Diffusion 2, Mistral-7B, Llama 3).[2]

3 Comparison with Previous Proposals

To understand the evolution of the AI Act, we compare it to prior proposals made during the AI Act legislative process. In particular, given our focus on foundation models and general-purpose AI, we consider the European Parliament's negotiated position from June 2023. This is the first formal EU position to introduce the subject of foundation models. In addition, we consider the proposal from Stanford researchers [Bommasani et al., 2023] aimed at supporting compromise during the legislative negotiation from December 1, 2023, six days prior to the Act's negotiation on December 7, 2023.

Takeaway 1: Different recipients of disclosures. All three texts center transparency, but the Parliament and Stanford proposals near-exclusively focus on transparency to the public. The AI Act, however, includes only one public-facing disclosure with all other transparency requirements being aimed towards either the government or downstream firms. As we discussed previously, this means the AI Act's contributions to advancing public understanding and AI accountability is likely to be weaker and less certain that what the Parliament and Stanford researchers envisioned.

Takeaway 2: Multi-tier approach. The Parliament proposal treats all providers equally, whereas the Stanford proposal and AI Act share the same structure of two tiers and a partial exemption for certain parties. The Stanford proposal recommends models being designated in the tier with more requirements if they have large demonstrated market impact, in line with the EU's approach in the Digital Services Act, where online platforms are subject to more scrutiny when they have at least 45 million EU monthly active users. The AI Act acknowledges measures of demonstrated impact as possible criteria the AI Office use, but instead default to using a compute-based threshold akin to the US [Executive Order 14110, 2023].

Takeaway 3: No independent researcher access. The Stanford proposal recommends guarantees that high-impact foundation models be accessible to independent researchers, given the demonstrated value of such research [Guha et al., 2023]. Such guarantees would emulate the EU Digital Services Act, which facilitates research of otherwise-opaque online platforms (e.g. to understand content moderation practices or the spread of disinformation) to advance the public interest. However, the AI Act does not contain such provisions, which is particularly relevant in light of recent calls from researchers for safe harbors to conduct research on issues and harms associated with foundation models [Longpre et al., 2024].

4 Conclusion

The EU AI Act is a seminal achievement in AI policy as the world's first comprehensive laws on AI. We analyze the Act's legal text to increase clarity as we await the Act's implementation, enforcement, and impact. We hope our work catalyzes greater scientific engagement in policymaking to yield superior evidence-based AI policy and better public outcomes.

[1] As of August 2024, the latest version is v0.0.8: see `https://opensource.org/deepdive/drafts`.

[2] In most cases, this is because these licenses impose use restrictions on the entities who can use the model or the purposes they use model for: these violate the OSI's criteria 5 and 6 about not discriminating on who can use the model and for what endeavors.

Acknowledgements We thank Sayash Kapoor, Shayne Longpre, Nestor Maslej, Betty Xiong, Daniel E. Ho, Marietje Schaake, Daniel Zhang, Tatsu Hashimoto, Connor Dunlop, Risto Uuk, Peter Cihon, Ben Brooks, Sabrina Kuespert, Kai Zenner and Luca Bertuzzi for valuable discussions related to the EU AI Act and the Stanford Institute for Human-Centered Artificial Intelligence (HAI) for supporting this work. RB was funded by the Stanford Lieberman Fellowship at the time of this work. This work was conducted prior to, and is entirely unrelated to, the involvement of RB in the EU AI Act Codes of Practice.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL `https://crfm.stanford.edu/assets/report.pdf`.

Epoch AI. Data on large-scale ai models, 2024. URL `https://epochai.org/data/large-scale-ai-models`.

Rishi Bommasani, Tatsunori Hashimoto, Daniel E. Ho, Marietje Schaake, and Percy Liang. Towards compromise: A concrete two-tier proposal for foundation models in the eu ai act, 2023. URL `https://crfm.stanford.edu/2023/12/01/ai-act-compromise.html`.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308, 2024. doi: 10.1126/science.adj0998. URL `https://www.science.org/doi/abs/10.1126/science.adj0998`.

Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The foundation model transparency index v1.1: May 2024, 2024. URL `https://arxiv.org/abs/2407.12929`.

Zuzanna Warso, Maximilian Gahntz, and Paul Keller. Sufficiently detailed?: A proposal for implementing the ai act's training data transparency requirement for gpai, 2024. URL `https://openfuture.eu/wp-content/uploads/2024/06/240618AIAtransparency_template_requirements-2.pdf`.

Executive Order 14110. Executive order on safe, secure, and trustworthy development and use of artificial intelligence, October 2023. URL `https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence`.

Group of Seven. Hiroshima process international code of conduct for organizations developing advanced ai syste, October 2023. URL `https://www.mofa.go.jp/files/100573473.pdf`.

Sara Hooker. On the limitations of compute thresholds as a governance strategy, 2024. URL `https://arxiv.org/abs/2407.05694`.

Rishi Bommasani. Drawing lines: Tiers for foundation models, 2023. URL https://crfm.stanford.edu/2023/11/18/tiers.html.

Yoshua Bengio, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, Shayne Longpre, et al. *International Scientific Report on the Safety of Advanced AI*. PhD thesis, Department for Science, Innovation and Technology, 2024.

Knut Blind, Mirko Böhm, Paula Grzegorzewska, Andrew Katz, Sachiko Muto, Sivan Pätsch, and Torben Schubert. The impact of open source software and hardware on technological independence, competitiveness and innovation in the eu economy. *Final Study Report. European Commission, Brussels, doi*, 10:430161, 2021.

Manuel Hoffmann, Frank Nagle, and Yanuo Zhou. The value of open source software. *Harvard Business School Strategy Unit Working Paper*, (24-038), 2024.

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen K Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. Position: On the societal impact of open foundation models. In *International Conference on Machine Learning*, pages 23082–23104. PMLR, 2024.

Jai Vipra and Anton Korinek. Market concentration implications of foundation models: The invisible hand of chatgpt. *The Brookings Institution*, 2023. URL https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt.

UK Competition and Markets Authority. Ai foundation models: Initial report, 2023. URL https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf.

US Federal Trade Commission. On open-weights foundation models, 2024. URL https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/open-weights-foundation-models.

US National Telecommunications and Information Administration. Report on dual-use foundation models with widely available model weights, 2024. URL https://www.ntia.gov/issues/artificial-intelligence/open-model-weights-report.

Neel Guha, Christie M. Lawrence, Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, and Daniel E. Ho. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review, Symposium on Legally Disruptive Emerging Technologies*, 2023.

Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Alex Pentland, Arvind Narayanan, Percy Liang, and Peter Henderson. Position: A safe harbor for AI evaluation and red teaming. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32691–32710. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/longpre24a.html.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The findings are tightly based on the coding and comparison with past proposals.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: The analysis in this work is tightly scoped, so we do not see any significant limitations to highlight, beyond indicating that, fundamentally, our ability to reason about the EU AI Act at this time is limited since we are understanding the legislative text rather than the Act's impact (because its requirements are not yet enforced at the time of writing).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theory involved.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: No experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: No code or data in traditional sense; the coding is included in the supplement.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the Code fo Ethics and do not see any matters that would constitute nonconformity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The primary societal benefit of this work is to increase public and scientific understanding, including across disciplinary boundaries, with respect to the EU AI Act. Given this is fairly self-evident, we do not discuss it at length. Beyond this, we do not foresee any significant negative societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not see any risks from the release of the coding.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The coding is available under a standard CC-BY license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The coding is released and is fairly self-explanatory with clear structure (e.g. column names, labeling scheme) that we also describe in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.