HUGGING FACE

Hugging Face Comments on BIS-2024-0047 / RIN 0694-AJ55: Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters

The proposed rule regarding *Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters* covers important aspects of reporting; it sets clear goals of who should be required to report and excludes small actors who might not have the resources to fulfill regular reporting requirements. It further recognizes the changing nature of AI technology by ensuring that requirements and thresholds can be updated. We offer recommendations to strengthen the reporting requirements based on our experience with model documentation and work on social impact evaluations. We have organized our feedback by the questions highlighted in the Request for Comments and the definitions included in the proposed rule.

About Hugging Face

Hugging Face is a community-oriented company based in the U.S. and France working to democratize good Machine Learning (ML), and has become the most widely used platform for sharing and collaborating on ML systems. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analyzing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI.

General Comment: Safeguarding Research to Support Safer Technology

Hugging Face is a platform that supports open science, collaboration on new models and AI technology, and replicable and transparent research into the properties of AI models. As such, we feel the need to draw attention to the specific conditions of open and collaborative research and development of AI to ensure that regulatory requirements remain compatible with those conditions. In particular, as regards the reporting requirements for models that meet the proposed threshold, we argue that they should be designed with the full range of stakeholders in mind, including smaller actors such as start-ups, non-profit, and academic developers who do play a critical role in enabling sustainable evidence-based policymaking on large models despite having access to fewer computational resources than larger companies.

Scientific replication efforts of models that were at the “frontier” of computational requirements at the time of their release in particular have been extremely valuable. Those efforts usually incur significantly decreased cost compared to the initial development due to several factors

including the decreasing cost of computational resourced over time, volunteer participation, increased reliance on public or donated data; and contrary to commercial development, those efforts do not depend on expensive inference cost for at-scale commercial deployment to meet their goals. Examples of such efforts include the BigScience project, which created BLOOM, the largest open-source multilingual AI model, through collaboration between research institutions, startups, and companies. This project, supported by a public grant to support the computational costs, replicated the capabilities of the then-"frontier" GPT-3 model a year after its release. BLOOM not only supports research on large language models but also plays a key role in advancing the understanding of safety, as it is widely accessible to researchers. Another example of large open-source AI models include Pythia, a suite for analyzing LLMs, and the open models developed by Allen AI, such as MOLMo, which has been shown to reach competitive performance to some of the most advanced commercial alternatives.

While such models may have limited direct economic impact, their value to research, innovation, and regulation is substantial. Among other contributions, these models have enabled research into vulnerabilities shared by commercial systems, including their susceptibility to jailbreaking, likelihood of memorizing and potentially leaking training information, and model extraction risks through API access. While efforts of these kinds have not yet trained a model involving more than 10^26 FlOps, we expect that access to sufficient public or private compute grants and funding will enable it in the future. Compliance with reporting requirements should therefore remain accessible to organizations with such unique needs as temporary coalitions of small actors or academic departments whose internal governance structure and legal representation differ from those of established commercial entities, especially in cases when the model development does not present consequent additional risk compared to existing commercial models. Additionally, much of the value of these efforts comes from their role in enabling **external** research on guardrails, evaluation, and more technically involved red-teaming, which complicates reporting of results. We make recommendations tailored for enabling compliance in those settings in the rest of this document.

Quarterly Notification Schedule

While we acknowledge the need for timely notification, the proposed quarterly notification as it applies to “any covered U.S. person”, including ones who are not currently developing any covered models, seems excessive. We propose instead that model developers should be required to report on covered activities when they are planning a training run and before the start of that training run, and that the reporting requirement of no applicable activity be removed, considering in particular the risks of administrative oversights in organizations with changing structures or personnel, including academic and non-profit organizations and start-ups in earlier growth stages.

Collection Thresholds

While the proposed collection thresholds based on computational operations and networking capabilities serve as a partial indicator of the resources required for reporting, they are not sufficient to adequately assess the risks associated with AI models. Compute thresholds, such as those outlined in E.O. 14110, offer insight into the scale of AI model training but fall short of capturing the broader risk landscape.

The risks posed by AI models cannot be understood in isolation from the context in which they are deployed, how they are used, and how accessible they are in terms of user interfaces and public access. Factors like the purpose of the model, the composition of its training data, its downstream applications, and its potential for misuse or harm are critical dimensions that go beyond raw computational power.

In order to reflect these considerations while preserving important open and collaborative research efforts, we propose a reporting scheme in two stages. We recommend that covered U.S. persons planning a training run that meets the given requirements should submit a broad description of their system that includes additional risk factors, including for example the inclusion of sources of data that contain information related to CBRN constraints, plans for deploying the model at scale, and plans for developing the model to enable entirely novel capabilities not available in existing widely available systems, but not additional information on system-level guardrails and testing occurring outside of the core model development activity (see our note on distinguishing definitions of AI **models** and AI **systems**). Models that do not present particular additional risk factors along any of those dimensions should be dispensed from further reporting. Models that do could be requested to provide additional information on testing and developers could remain available for follow-up outreach.

This approach aligns with frameworks like the EU AI Act, where developers who meet a compute-based risk threshold are able to argue that their model does not present systemic risk if it has comparable capabilities to existing models. It also ensures that attention is focused on the models with the highest risk, while allowing open and collaborative research to keep pace with commercial development and provide an important scientific evidence basis to support further regulation.

Furthermore, the inventory of models generated from this reporting could be coordinated with other disclosure processes, including for example systems that allow experts and users of AI systems to follow a process of coordinated disclosures when flaws or risks related to AI models and systems are identified.

Definitions

AI red-teaming

The current definition of *AI red-teaming* could benefit from greater clarity and depth, particularly around key areas like the scope of the red-teaming, how experts are selected, and who gets to contribute to the overall safety and reliability testing of AI systems.

Safety testing should be applied to both the AI model as well as possible guardrails (see c(3)(ii)) to ensure knowledge about both the model capabilities and the guardrails employed on the model. Those guard rails can change faster than retraining a model and research has shown that these guard rails are not immune to jailbreaking.

These developments are most often publicly reported by external researchers. The current definition of AI red-teaming requires "*collaboration with developers of AI*". It should be clarified that the involvement of third party actors, who are not involved in the development of the AI system, is desirable. Encouraging open, collaborative approaches would allow diverse voices to contribute to defining risk thresholds and management strategies, fostering a more comprehensive risk management framework. Establishing guidelines for safe public and expert participation in red-teaming, along with safe harbor clauses, would expand its reach and effectiveness. Additionally, organizations would benefit from anonymized, shared findings, enabling collective understanding of emerging threats. This would foster a culture of shared responsibility, where the entire AI community contributes to safer, more reliable systems.

While red-teaming is helpful for identifying flaws and vulnerabilities in AI systems, such as harmful outputs or unforeseen behaviors, it should be part of a broader, collaborative risk management strategy, integrating social impact evaluations.

Any model evaluation framework needs to ensure that the testing of models is not replicating what is part of the training data, i.e., data leakage. A level of reporting on the training data is required to ensure that evaluation results are reflecting the required standards.

AI model and **AI system**

The current definitions of an AI model and system are currently too close to support regulatory efforts that have to rely on important distinctions between both. We recommend strengthening that distinction, taking inspiration for example from definitions of texts like the EU AI Act that make different requirements of models and systems.

An AI model is best understood as information: typically encoded as a set of model weights that represent statistics extracted from training data. An AI model cannot, by itself, produce outputs given inputs without additional software and hardware components.

An AI system is obtained by combining one or several models with different software components to process inputs and outputs and run information through the model weights, as well as with hardware (typically a *Large-scale computing cluster*) to enable the required computation.

These differences are meaningful on several accounts. Firstly, guardrails are often integrated in a system as input or output processing outside of the core models, and are better discussed as a property of the system than of the models; in particular, model-level safety-finetuning type approaches have been shown to be brittle both for commercial APIs and for models with widely available weights. Secondly, the cost of deployment of a model is a significant factor in assessing the risks posed by its deployment. A model that is easily accessible to hundreds of millions of users through commercial deployment supported by a large cluster may pose risks on a different scale than a model released as a set of weights.

Dual-use foundation model

The definition of *dual-use foundation model* contains a requirement of size in parameters ("*(c) Contains at least tens of billions of parameters*"), which might be outdated as model sizes are changing over time. A definition of *dual-use foundation model* should focus on the application and context of the model rather than its size.