Convergence Analysis' Official comment on Proposed Rule by the Industry and Security Bureau on 09/11/2024:

Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters.

Convergence Analysis is grateful for the opportunity to comment on the proposed rule "Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters" (Docket No. 240905-0231). We applaud BIS's timely efforts to implement reporting requirements as mandated by Executive Order 14110.

Convergence Analysis is a United States-registered non-profit AI safety think tank. We recently published an in-depth report, **AI Model Registries: A Foundational Tool for AI Governance**, in collaboration with Gillian Hadfield (University of Toronto), Tino Cuellar (President of Carnegie), and Tim O'Reilly (O'Reilly Media), based on their op-ed proposing a national registry for AI models published last year.

The report, which we will refer to below as 'our report,' was developed in consultation with staff from RAND, OpenAI, CLTR and the Centre for AI Governance.

Based on our report, we offer the following comments on the BIS's Proposed Rule:

1. More Specific Collection Thresholds for Dual-Use Frontier Models

Here we propose and address three concerns with the proposed thresholds for sending questions to covered US persons. Our recommendations do not encompass the Proposed Rule's parameters for large-scale compute clusters.

a. Compute alone is not a perfect proxy for model capability

The Proposed Rule currently defines 'applicable activities' in terms of compute. While compute is a useful proxy for model capability, it is insufficient on its own. Research into scaling laws has demonstrated that model capabilities depend on a balance of compute, training data quality and quantity, and model size. The architecture of the model is also a powerful modifier. Given that data on each variable contextualizes the others, we recommend expanding on the template laid out in Executive Order 14110 by including a combination of high level information on the following three categories:

- **Model architecture:** The architecture of a model provides important context for proxies of capability. For example, what are called *sparse* or *sparsely activated* models[1] often have many times more parameters than non-sparse models with equivalent capability, as only a small fraction of the parameters in sparse models are active during operation[2]. Reporting a high-level description of model architecture would provide greater regulatory insight and important context, while still protecting proprietary information and allaying security concerns.
    - Developers should provide a high-level technical description of the model architecture, sufficient for an expert in the field to distinguish it from similar models with different performance or functions.
    - The description should include the general type of architecture (e.g., transformer, mixture-of-experts, etc.) and any significant innovations or departures from standard architectures.
    - Developers should report the number of layers and the types of layers used (e.g. attention layers, feed-forward layers) without disclosing precise configurations.
    - Developers should disclose if the model uses any form of external memory or knowledge retrieval systems.
    - The description should include information on whether the model uses multi-modal inputs or outputs, specifying the types of data it can process (e.g., text, images, audio). These descriptions should not be so detailed as to allow replication of the model or to reveal trade secrets that could significantly advantage competitors. The reporting system should also include provisions for periodic reviews of architectural disclosure requirements to ensure they remain relevant and effective as AI technology evolves

- **Model size**: We recommend that developers should report a measurement of the total number of parameters for non-sparse models, or the average number of active parameters during use across a wide range of model inputs for sparsely activated models. To account for measurement uncertainty, the total number of parameters for a model should be accurate to within 10% of the true value.

- **Amounts and types of training data:** We recommend that the amount and types of data used to train a model should be reported. The amount should be measured by the number of tokens, and be accurate to within 5% of the correct value to account for difficulties in measurement. Developers should be required to report the type of data by selecting categories from a list, for example whether any of the following were used:
    - Text
    - Images, including subcategories such as labeled images of people
    - Audio, including subcategories such as isolated audio of human voices
    - Video
    - Genetic, biological, or bioinformatics data

[1] Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity - Fedus et al.

[2] A Review of Sparse Expert Models in Deep Learning - Fedus et al.

- ○ Toxicity, volatility, etc. of chemicals or biological products

This information would be security relevant, non-hazardous, and not overly burdensome for AI labs to provide. For more detailed discussions of what information could be required, see our report on model registries under the following sections: *What thresholds should a model exceed to qualify for inclusion?; Key concept: scaling laws; Training data; Model size and parameters; Compute used for training; Model architecture*.

b. Focusing on training runs only may allow new capabilities to be developed without being reported.

If 'applicable activities' only covers training runs, reporting may fail to capture new capabilities. Models can often be fine-tuned or adapted using much less compute than their initial training. It is feasible that a model could be iteratively improved by methods other than compute-intensive training, such as reinforcement learning from human feedback and fine-tuning, to the point where its performance may be notably different. For example, GPT3.5 was trained in March 2022, but was further fine-tuned and released to the public as ChatGPT in November[3]. This is an important change in capabilities and access, but did not involve a large training run and thus could be missed in the current proposed reporting regime. Indeed, many models in the GPT-3 family, of which there are now over fifty[4], were not independently trained but rather fine-tuned or otherwise adjusted for specific purposes. In this case the fine-tuning was benign, but it's plausible that fine-tuning could make a model more likely to provide hazardous information. For example, fine-tuning could affect a model's "willingness" to disclose information that could lower the barrier of entry to building biological or chemical weapons[5].

In our report, we advocate for a system where developers can and must provide updates to the government on such changes to models, as opposed to making quarterly reports as described in the proposed rule (we discuss our proposed alternative in more detail below under *2. Pragmatic Reporting Schedule*). We recommend that developers may be allowed to add and update qualifying models as a new model version as part of an existing model family. Such a model family would primarily record the entire set of reporting requirements for the most capable model version along each key measurable dimension.

We recommend that developers only be required to submit a complete report for a model version in the case that the new model exceeds the most capable model versions in its family by a certain amount (e.g. 20% of model size), or released some time interval after initially reported (e.g. after 3 years of initial report). If it does not exceed the most capable model versions and was deployed within the time interval of the initial report, it will be sufficient to simply report the

[3] "ChatGPT: Optimizing Language Models for Dialogue". *OpenAI*. November 30, 2022. Archived from the original on November 30, 2022. Retrieved January 13, 2023.

[4] The GPT-3 Family: 50+ Models - Alan D Thompson

[5] Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools - Jonas Sandbrink

name of the new version for tracking purposes. Ideally, these key measurable dimensions would be based on capability assessments. As there is no existing consensus on sufficient assessments of capability, we recommend currently using the following key measurable dimensions as proxies for the “most capable” model, also discussed above in 1.a:

1. Model size
2. Total compute used during training and retraining
3. Amount of training data
4. Specific powerful capabilities, such as the ability to generate CBRN infohazards, generate deepfakes, conduct autonomous replication, or improve cyberwarfare abilities.

For the first three key measurable dimensions described, a new report would be required when a new version exceeds the previously most capable model by 20%. For the measurable dimension of “specific powerful capabilities”, a new submission would be required when a new version exceeds the previous most capable version in its results from a related capability evaluation by a certain amount (e.g. 20%), or when crossing a certain threshold score (e.g. scoring 80% in an autonomous replication capability eval). Note, however, that reliable evaluations of this sort do not yet exist.

We also propose that models must be reported as a new family if they’re deployed more than 2 years after initially reported. This is important, as algorithmic progress means that models in 2030 could be far more capable than models in 2027 while still the same size and trained with the same compute and training data.

.

c. Focusing on training runs may allow new safety-relevant changes, such as model security and evaluation results, to go unreported.

Beyond model capabilities, other security-relevant information may be changed without qualifying as ‘applicable activities.’ For example, changes in cyber or physical security of model weights would not be reported, despite having potentially significant national security implications. New evaluations may uncover previously undisclosed model capabilities, with no obligation to report these capabilities until the next training run. For example, large training runs on the GPT series at OpenAI have been spaced out over many months and even up to two years[6]. In this way, information pertaining to the safety and reliability of dual-use foundation models may be excluded from reports as long as no new significant training has taken place.

We recommend that ‘applicable activities’ be expanded to include specific information pertaining to the safety and reliability of dual-use foundation models that are already subject to reporting requirements, including any changes to model physical and cyber security and new model evaluation results. We provide a list of recommended additional questions and specific

[6] Our World In Data - Computation used to train notable artificial intelligence systems, by domain (Restricted to GPT series)

information to be reported under *4. Suggestions on Specific Questions to ask Covered Persons upon Notification* below.

It may be difficult to prescribe specific thresholds for what changes qualify as additional ‘applicable activities’, so an alternative is to require regular updates from providers of covered models on this information, regardless of whether information from the last report has changed. We discuss this approach more in the Pragmatic Reporting Schedule section below.

2. Pragmatic Reporting Schedule

Requiring quarterly reports is likely to provoke resistance to the Proposed Rule from AI developers. Based on our discussions with staff at large AI labs, quarterly reports on current and future activities could create constant administrative pressure on labs, which will increase as more training runs begin to qualify as ‘applicable activities’.

We also raise concerns that the proposed definition of ‘applicable activities’ risks overlooking significant changes in model capabilities models (1.b), and significant changes to model security or evaluation results (1.c).

As an alternative to the quarterly updates suggested in the Proposed Rule, our report recommends a system that requires updates directly based on new model capabilities to address these concerns. As we discuss above in 1.b, we recommend that developers update BIS on model changes as they occur, rather than quarterly. New models can be added to existing model families without full reports if they don't significantly exceed capabilities, and full reports are only required for models exceeding family capabilities by 20% or after 2 years. Capability thresholds are determined using the proxies identified above in section 1.a. This system ensures significant advancements are properly documented, while reducing the burden of reporting for both BIS and developers.

To ensure that information remains accurate, developers should be required to update their reports twice annually, regardless of updates to the model family. This should include information on model security and model evaluation results, addressing concerns raised in 1.c.

This system would reduce the overhead of reporting updates for developers, by concentrating reporting requirements solely on the most capable model versions. New model development could occur without necessitating a report, as long as they do not meaningfully exceed the most capable version along a key measurable dimension.

For more details on our proposal, see sections *What Should Qualify for Inclusion on the Registry?*

3. Secure Data collection and Storage

We recommend BIS carefully consider the utility and sensitivity of different types of AI model information in relation to its regulatory goals. Key objectives, such as identifying dual-use systems and assessing adversarial risks, may be achieved using less sensitive data. For instance, high-level capability assessments, standardized evaluation metrics, and conformity with cyber and physical security standards can provide valuable insights without exposing vulnerabilities or sensitive details.

However, we acknowledge that some sensitive information might be necessary for specific regulatory purposes. In these cases, we suggest implementing stringent security measures and storing such data separately from general reporting information. The key is to recognize that not all useful data is equally sensitive, and not all sensitive data is equally useful for every regulatory goal.

This nuanced approach would allow BIS to tailor its data collection strategy, balancing effective oversight with minimized security risks. By differentiating between types of information and their specific utilities, BIS can make informed decisions about what data to collect and how to protect it, rather than adopting a one-size-fits-all approach to information gathering and storage.

4. Suggestions on Specific Questions to Ask Covered Persons upon Notification

The Proposed Rule adopts broad topic specifications for questions to covered US persons from Executive Order 14110. To enhance the Rule's effectiveness, we propose additional, specific topics to ensure comprehensive awareness of existing dual-use foundation models, create visibility into their vulnerabilities to malfunction or unauthorized access, and understand the implications of their potential use by adversaries. These focused topics aim to better equip BIS with the information necessary to address critical security concerns in AI development and deployment.

a. Open-Source Status

Open models pose unique regulatory challenges compared to closed-source models, as they cannot be easily controlled once deployed and are more vulnerable to misuse and replication. This may necessitate stricter pre-deployment safety assessments and targeted regulations. Additionally, open-sourcing can accelerate AI capability development, potentially widening the gap between AI capabilities and safety measures. Therefore, developers should be asked what license their models are deployed under in addition to ownership and possession of model weights.

b. Functions of the Model

The purpose and intended use of the models provides important context to the government, improving understanding of potential real-world applications and associated risks. However, defining use cases for AI can be complex, especially for versatile models like LLMs, which may have capabilities beyond their initial training objectives. To address this complexity, developers could be required to report basic explanations of primary uses, examples of potential alternative uses identified through safety assessments and benchmarks, and links to model documentation such as research papers, API references, and user guides. This approach would offer valuable insights without placing an unreasonable burden on developers, aligning with practices in other industries while accommodating the unique characteristics of AI systems.

c. Post-Deployment Monitoring practices

Post-deployment monitoring of AI systems is crucial due to the uncertainty and complexity inherent in AI behavior. Indeed, many AI labs already engage in various monitoring practices. Developers should be asked to report their monitoring practice, including safety KPIs, response thresholds, incident protocols, and policies for reviewing these practices. This information would improve coordination and transparency without imposing excessive regulatory burdens. It would help the BIS ensure that emerging capabilities, use-cases, or incidents with national security implications will be recognised in time to respond.

d. Proxies for model capability

Proxies for model capability are important as discussed in part 1.a of this comment. Specifically, developers should be asked to report:

- Model Size & Parameters
- Training Data
- Model Architecture

These could be seen as distinct topics, or simply provide guidance for developing questions around 702.7(b)(2)(iv) ‘Other information pertaining to the safety and reliability of dual-use foundation models, or activities or risks that present concerns to U.S. national security.’

Contact

For more information, please contact us at policy@convergenceanalysis.org or visit our website at https://www.convergenceanalysis.org/.