ALM | LAW.COM

New York Law Journal

TECHNOLOGY TODAY

ARTIFICIAL INTELLIGENCE

The Knowns and Unknowns of Crossing the AI Frontier

By Katherine B. Forrest
July 29, 2024

One of the ways we can tell that technological developments in AI are moving fast— really fast— is the current dialogue relating to AI "Frontier" models. A Frontier model is a "highly capable model" that "could possess capabilities sufficient to pose severe risks to public safety." (Anderljung, et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," November, 2023).

The White House Executive Order ("E.O.") on AI, issued on Oct. 30, 2023, refers to these models as "dual-use" foundation models. The U.K. Government Office for Science has published a special report on "Future Risks of Frontier AI," and the AI Seoul Summit, held in May 2024, was followed by an "International Scientific Report on the Safety of Advanced AI."

KATHERINE B. FORREST *is a partner in Paul, Weiss, Rifkind, Wharton & Garrison's litigation department and a member of the antitrust practice group. She previously served as a U.S. District Judge for the Southern District of New York and as a Deputy Assistant Attorney General in the U.S. Department of Justice's antitrust division.*

Katherine B. Forrest

There is broad agreement that the capabilities of large scale neural networks such as those that enable the generative AI foundation models that the world woke up to with the issuance of ChatGPT in the late fall of 2022, are rapidly increasing.

Models trained on huge data sets that have billions of "parameters," and can perform a broad range of tasks unsupervised, can approach or cross the threshold into a frontier model. Parameters are the values assigned to the data the neural network is trained on and establish relationships between pieces of data. As additional data is used to train a model, the parameters can adjust.

AI frontier models are characterized by their ability to be used for both good and bad purposes—hence the term "dual-use" model in the E.O. On the one hand they can provide the basis for extraordinary advances in pharma, content

ADOBE STOCK

creation, and numerous other domains. What distinguishes these models from less capable ones are characteristics such as enhanced memory, the ability to plan and reason, higher rates of accuracy and reduced hallucinations, an ability to function autonomously or semi-autonomously, and some ability to engage in self-improvement.

We already know that complex generative AI models available today have displayed unexpected "emergent" abilities. That is, abilities that "emerged" over time and were not apparent or known to exist during the design or training process. Many emergent capabilities—including various forms of reasoning, mathematical abilities and self-teaching—were discovered after models were released for public use.

The millions and millions of humans that started using and experimenting with these models discovered that they could do far more than even their developers had expected. None of us know just what emergent capabilities users will discover when they start using frontier models. The known unknown of emergent capabilities in Frontier models is only one of regulators' concerns.

Among the issues that have given rise to the broad safety debate are that such models, in the wrong hands, will enable those without (for instance) specialized knowledge in chemistry, biology or physics, to create chemical, biological, radiological or even nuclear (CBRN) events.

Additional concerns include an ability for such models to create and proliferate disinformation at levels heretofore unseen, or to find cyber security vulnerabilities. But even beyond that, there is a concern that some of these models might have ways of evading human control through deception and misdirection. In other words, in the wrong hands, frontier models could theoretically find ways of evading human control while doing serious and even catastrophic damage.

The known unknown of emergent capabilities in Frontier models is only one of regulators' concerns.

While we haven't seen these highly negative scenarios yet, regulators worldwide are trying to find ways to get ahead of what they see as a problem that will occur, even if we are not quite sure of how or when.

The United States has tasked a variety of agencies to propose regulations directed at public health, safety and national security risks. In an effort to further define these models in more than simply a way that references broad capabilities, both the E.U. A.I. Act and the E.O. have set forth a quantitative measure based on the number of "floating point operations" or FLOPs. A FLOP is, in reductionist terms, how many mathematical operations a computer or neural network can perform every second. The E.U. has posited a threshold of 10^{25} FLOPs; the

White House E.O. has increased that to 10^{26} (but there is some discussion of lowering the U.S. threshold.)

For companies that are experimenting with highly capable models to accomplish useful and beneficial tasks, perhaps models that will be able to create significant medical breakthroughs, they need to carefully monitor regulatory developments in this area.

Among the proposed regulatory obligations would be model registration, the need to report on training, development and production, a requirement to provide transparency with regard to testing and mitigation measures if necessary, the need to provide proof of adequate cyber security measures, verification of shut down capabilities, and a version of "know your customer" (or "KYC") information. Some states in the U.S. (e.g. California S.B. 1047), are seeking to pass similar as well as variations on such requirements.

As of today, the U.S. Department of Commerce has been tasked with establishing specific regulations. It has sought information from stakeholders, but has not yet promulgated final rule-making. The Supreme Court's recent decision that eliminates *Chevron* deference, *Loper Bright Enterprises v. Raimondo*, may throw additional uncertainty into this area if such regulations (once passed), are challenged.

In addition to rule-making that will regulate Frontier models deployed domestically, the U.S. has imposed certain export controls on some components of AI systems such as semiconductors. The Committee on Foreign Investment in the United States ("CFIUS"), an interagency committee of the federal government that reviews implications of certain foreign investments in U.S. businesses, has rules that encompass critical technology and sensitive personal data that can implicate AI systems under certain circumstances. The Department of Treasury also restricts certain outbound investments in AI.

There are a number of open questions with all regulation of AI, and in particular with regard to Frontier models. Over the next year or more, there will be additional clarity (I predict) around what capabilities and quantification measures constitute a Frontier model, how to measure those metrics, and what kind of controls should or need to be imposed.

An open question that may not be resolved in that same timeframe concerns the tools that will be most effective at implementing the desired controls that can provide the best mitigation measures. Lawyers counseling clients in the area of Frontier models have their work cut out for them. The keys are to be aware of a quickly changing regulatory landscape with multijurisdictional implications, and to be nimble.

Katherine B. Forrest *is a partner in Paul, Weiss, Rifkind, Wharton & Garrison's litigation department and a member of the antitrust practice group. She previously served as a U.S. District Judge for the Southern District of New York and as a Deputy Assistant Attorney General in the U.S. Department of Justice's antitrust division.*