Disclaimer

The responses to this Request for Information (RFI) reflect the opinions of the authors. They are not necessarily the opinions of the organizations that employ the authors. Readers shall not construe these comments as reflective of current policy or recommended future policy by federal, state, or local governments. This is not a position paper from any federal, state, or local government, from contractors, or from the industry.

American Council for Technology-Industry Advisory Council (ACT-IAC)

The American Council for Technology-Industry Advisory Council (ACT-IAC) is a non-profit educational organization established to accelerate government mission outcomes through collaboration, leadership, and education. ACT-IAC provides a unique, objective, and trusted forum where government and industry executives are working together to improve public services and agency operations through the use of technology. ACT-IAC contributes to better communication between government and industry, collaborative and innovative problem solving, and a more professional and qualified workforce.

The information, conclusions, and recommendations contained in this publication were produced by volunteers from government and industry who share the ACT-IAC vision of a more effective and innovative government. ACT-IAC volunteers represent a wide diversity of organizations (public and private) and functions. These volunteers use the ACT-IAC collaborative process, refined over forty years of experience, to produce outcomes that are consensus-based.

To maintain the objectivity and integrity of its collaborative process, ACT-IAC welcomes the participation of all public and private organizations committed to improving the delivery of public services through the effective and efficient use of technology. For additional information, visit the ACT-IAC website at www.actiac.org.

Emerging Technology Community of Interest

The Emerging Technology COI AI Working Group collaborates with Federal CXOs and other government executives responsible for identifying, assessing, and deploying emerging technology and maturing it to become a major component of the IT and business strategy, as well as industry, government, academia, and the greater community to provide products, services, processes, and business models enabling innovative approaches for solving government issues and challenges.

Artificial Intelligence Working Group

The Artificial Intelligence Working Group (AIWG) engages with subject matter experts in the areas of artificial intelligence/machine learning (AI/ML), advanced analytics, statistics, and other relevant areas from across government and industry to inform and advise government agencies and the federal contractors. The AIWG provides learned inputs and knowledge resources that organizations need to 1) Identify areas that would benefit from AI/ML implementation, 2) Optimize AI/ML implementations, maintenance, and management, and 3) Provide a common space to share best practices that drive mission and operational value.

Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights

1. How should NTIA define "open" or "widely available" when thinking about foundation models and model weights?

When thinking about foundational models and model weights, the NTIA should define "open" or "widely available" as offering broad access to its inner workings, including the model's code, comments, and associated weights, all publicly available in a clear text format. This transparency will foster trust and enable public scrutiny of the model's functionality.

It is important to acknowledge that there may be limitations to complete openness due to intellectual property considerations, and a spectrum of openness might be appropriate depending on the specific model and its potential risks and benefits. For instance, some models might require stricter controls due to safety concerns.

2. How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?

Zoë Brammer's six categories of risk (in "How Does Access Impact Risk")[1] provide a useful framework for this discussion:

- *Race to the bottom:* Open models mitigate cutting corners by providing transparency. The risks here may be analogous to cryptography, which spawned massive scams and pyramid schemes alongside legitimate endeavors.
- *Malicious use:* Malicious actors continuously seek to exploit closed systems; making weights and models open eliminates any barrier to such work. The more open the model, the sooner downstream users will exploit it.
- *Capability overhang:* As with malicious use, the risk of emergent capabilities is accelerated by openness. On the other hand, open models mitigate this, providing the ability for unaffiliated researchers to work backwards from emergent behavior to its sources.
- *Compliance Failure:* Where anyone can tinker with a model, regulation becomes a polite fiction. Consider 3-D printing, or streaming music. Open models virtually ensure that bad actors will ignore any legal or technical safeguards that governments or industry attempt to enforce.
- *Human out of the Loop:* Brammer argues that the risk humans might lose control of AI systems is non-trivial, especially as automating human tasks is one of the stated goals of industry. Here there are a combination of factors to consider with regards to risk. Open models will spawn iterations that are trained for specific uses and reduce the ability of governments and industry to regulate outcomes. Strict oversight and regulation of closed models with the goal that open models be tested and "well-understood" by federal and/or industry agents would mitigate this risk. Market forces work against this kind of limitation.
- *Reinforcing bias:* Iterative model development will amplify biases and flaws from poorly understood source models. In this context, open models allow non-affiliated researchers to understand built in biases, in a similar manner to how citations in academic papers allow readers some understanding of the genesis and development of arguments.

In conclusion, while risk is associated with both public and non-public model weights, the lack of transparency associated with closed models presents a significant challenge. While there are clearly instances where closed models are appropriate, public scrutiny of model weights allows for a more informed assessment of potential risks, fostering trust and improving overall safety. Without such openness, the public is left to rely solely on assurances from developers, a strategy increasingly insufficient in today's AI landscape.

[1] Brammer, Zoe: *How Does Access Impact Risk*, p. 13-14

b. Could open foundation models reduce equity in rights and safety-impacting AI systems (e.g., healthcare, education, criminal justice, housing, online platforms, etc.)?

We find the premise of this question problematic. Absent oversight and an understanding of the inputs used to train the model, any inequities present will be perpetuated, regardless if the model is open or closed. Theoretically, open foundation models would reduce inequity in rights and safety-impacting AI systems because open foundation models enable greater oversight by parties with vested interests and the power to enforce/influence regulations.

c. What, if any, risks related to privacy could result from the wide availability of model weights?

Any model which contains material that can be used to violate privacy either through direct input or through the aggregation of public information that was previously difficult to access will result in negative outcomes regarding privacy. In specific regards to "numerical parameter[s] within an AI model that help [. . .] determine the model's output in response to inputs," such risks seem minimal. The hazard lies in the data used to train the model, not in the manner which models process data - what risk there is would be in malicious actors' ability to work backwards from outcomes to training data. Having open and available model weights would create a "level playing field" for the users of the product and reduce the risk of unintended exposure in a manner similar to open-source software.

d. Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy?

Knowing the algorithm and the weight models could allow nefarious actors to accomplish their goal, which could have significant security implications. Such risks could involve the healthcare sector, defense, business and safety.

i. How do these risks compare to those associated with closed models?

With closed models, replication would be more challenging as the model weights and the algorithm would not be known. A bad actor could make assumptions, which can lead to a negative outcome.

e. What, if any, risks could result from differences in access to widely available models across different jurisdictions?

The risks are mainly associated with the nature of the AI tool and its desired outcome. Access to the model by a bad actor can lead to unsafe and bad results. AI tools need to be closely managed. Management of the AI tools (governance) will lead to different outcomes as each jurisdiction manages them with more or less available resources and experience.

f. Which are the most severe, and which are the most likely risks described in answering the questions above? How do these sets of risks relate to each other, if at all?

The most critical risk lies in the potential for harm, even death, if these tools are not implemented and managed responsibly. The most likely risk to emerge in the scramble for market dominance is the proliferation of inequitable data models that perpetuate unequal outcomes. The race to market share further exacerbates these risks. Established players might prioritize secrecy to maintain their edge, while challengers focus on cost reduction to gain a foothold. In this scramble, both sides are likely to neglect crucial safeguards, such as ensuring equitable and unbiased outcomes. Lives and property are placed at risk when these powerful tools are built on unreliable data and operate without ethical considerations. Developing and deploying AI responsibly requires prioritizing safety, transparency, and fairness throughout the process.

Concerns about the quality and limitations of training data span the risks outlined above. AI models learn from the data they are fed, and any biases or inaccuracies present in that data will be reflected in the model's outputs. If the data used for training doesn't accurately reflect the real-world situations the AI will encounter, the results can be disastrous. For instance, an AI trained on a limited geographical dataset might malfunction when deployed in a new location. Similarly, a model blind to biases in its training data may perpetuate those biases in its decisions, leading to unfair and potentially harmful outcomes. Open model weights increase the opportunity for understanding bias in systems, which is vital to ensuring positive, safe outcomes.

3. What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?

Open foundation models and foundation model weights signify a transformative shift in the landscape of artificial intelligence (AI), presenting notable advantages over closed models in terms of innovation, transparency and security, and fostering equitable AI governance. In today's digital era, where AI holds substantial sway across diverse sectors, comprehending the societal implications of different model governance approaches is imperative.

Serving as engines for innovation, open foundation models drive progress in AI research and application development. Their accessibility and adaptability facilitate collaborative endeavors among researchers, industry experts, and policymakers, fostering breakthroughs in areas such as interpretability, security, and efficiency techniques. The transparency and openness inherent in these models encourage knowledge-sharing and experimentation, accelerating the pace of technological advancement and broadening the horizons of AI capabilities.

Open foundation models prioritize transparency as a cornerstone principle in AI development and deployment, encompassing aspects of security. Unlike closed models, which often operate covertly, open models prioritize transparency by providing visibility into the development process and downstream impacts. This commitment to transparency not only enhances accountability and trust but also strengthens security measures. By openly sharing information about model architecture, training data, and decision-making processes, open foundation models empower stakeholders to assess potential security vulnerabilities, identify risks, and implement robust safeguards. This proactive approach to transparency in AI development contributes to bolstering cybersecurity defenses, mitigating the risks associated with malicious exploitation or unauthorized access to AI systems. Therefore, transparency plays a pivotal role in enhancing both the integrity and security of AI systems, aligning with broader objectives of ethical AI governance and responsible innovation.

Furthermore, open foundation models redefine AI governance dynamics by democratizing decision-making processes and dispersing control among a wider array of stakeholders. In contrast to closed models, which often centralize authority within a select group of developers and stakeholders, open models empower downstream users to actively shape the trajectory and applications of AI technologies. This decentralized paradigm not only mitigates the risks linked with monopolistic control but also fosters a culture of collaboration and inclusivity within the AI ecosystem, thereby cultivating a sense of ownership and accountability.

In conclusion, open foundation models and foundation model weights offer distinct advantages over closed models by driving innovation, ensuring transparency, and fostering equitable AI governance in AI development and deployment. By democratizing decision-making processes and fostering collaboration, these models enable stakeholders to actively contribute to shaping the future of AI technologies. Embracing the adoption and governance of open foundation models and weights is essential for realizing the full potential of AI while addressing ethical, social, and economic considerations in an increasingly digitized world.

7. c. When, if ever, should entities deploying AI disclose to users or the general public that they are using open foundation models either with or without widely available weights?

Entities that deploy AI should ALWAYS disclose to users and/or the general public such use of AI whether or not it involves open or closed models and model weights. Similar to how entities are required to disclose the use of credit checks and background checks/investigations. In particular, those who are impacted by the output or use of AI should be made aware. Such disclosures will assist with the adoption of the AI tools; better adoption usually takes place when the algorithms are made available to stakeholders.

8. In the face of continually changing technology, and given unforeseen risks and benefits, how can governments, companies, and individuals make decisions or plans today about open foundation models that will be useful in the future?

The key to making decisions and plans about future opportunities is to remain vigilant to the state of the art and to participate (to the extent possible) in the conversations as they develop. To this end there is a need for an effective Governance Structure around developing technologies that can set goals, supply resources to assess risk and to provide sufficient implementation guidance. There should be continuous monitoring of AI tools for effectiveness and safety so that when issues arise, they would be promptly addressed.

For governments, standards must be established and continually evaluated for relevance, lest market forces outstrip their ability to adapt. Industry will need to be part of the regulatory dialogue and early engagement will help to drive outcomes that are based on deep subject matter experience. Individuals will remain at the mercy of larger forces that determine the direction and development of open foundations models, but can advocate for greater transparency through legislation and organization in response to markets.

9. What other issues, topics, or adjacent technological advancements should we consider when analyzing risks and benefits of dual-use foundation models with widely available model weights?

While considering a risk/benefit analysis of the potential of dual-use foundation models, we prioritize risk mitigation. Continued investment in privacy and robust cybersecurity practices is essential. Understanding stakeholder needs and potential biases within the models themselves is crucial for fair and equitable outcomes. Finally, establishing clear lines of accountability and exploring potential regulations, similar to how guardrails were needed for NFTs, will help ensure responsible use of these powerful tools.

Contributors

Dr. Christopher Martin
Eric Fears, MS-EECS
Thompson Boyd
Alec McLean