Center for AI Policy

To: National Telecommunications and Information Administration
Author: Thomas Larsen, Center for AI Policy
RE: Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights

Executive Summary

There has been substantial debate over whether open-sourcing artificial intelligence (AI) is good for the world. As is often the case in AI, we think the answer depends on which AI you're talking about. For the majority of AI models, we believe that open development and release is beneficial for society. **However, releasing highly capable AI model weights would substantially elevate national security risks and hence should not be allowed.**

We think that there are two types of systems that should be considered high-risk and therefore should not be allowed to be open-sourced:

1. **WMD-like AI**: AI systems that enable bad actors to develop WMDs much more effectively.
2. **Artificial General Intelligence (AGI)**: AI systems that are as generally competent as humans across a wide range of tasks.

Open weight high risk AI systems are dangerous because a large number of actors would gain access to very destructive technology. Fortunately, current AI systems are probably not high risk[1]. There remains a great deal of uncertainty about the length of time until we develop high risk AI systems: Shane Legg, the Co-Founder of DeepMind, predicts that a 50/50 chance of AGI by 2028, but many other experts think that AGI timelines are substantially longer. Dario Amodei, the CEO of Anthropic, predicted that AIs will be able to build highly dangerous bioweapons within 2 years, suggesting that WMD-like AI may be just around the corner.

In this response, we first outline key benefits and risks of widely available model weights. Then, we discuss policy recommendations for mitigating threats. Our main recommendations are to develop better monitoring and forecasting to predict when advanced AI might be developed, and to prevent high-risk AI models from being open sourced. Finally, we respond to selected questions from the NTIA Request For Comment.

[1] Unfortunately, our capabilities evaluation methodology is imperfect, and so we cannot be completely certain that the current models are not high risk – this is discussed in the capabilities elicitation section.

Benefits of widely available model weights

This section addresses the NTIA RFC Question 3: What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?

Accelerate AI safety research. Open model weights accelerate a large number of AI safety research agendas. Most interpretability research is done on large language models with widely available model weights. This especially helps smaller labs and academics have access to models that are closer to the cutting edge of capabilities.

Accelerate the adoption of positive AI use cases. There are many positive uses of AI systems – AI can be used to: increase productivity, improve decision making, and accelerate economic growth. Increasing access to model weights accelerates that research.

Enable Transparency. Widely available model weights enable more thorough understanding of those models, and enables red teaming efforts to identify potential vulnerabilities. Historically, open source software has been more secure, as the broader community can identify and propose mitigations for risks. This carries over somewhat to model weights because, as with other software, friendly actors can identify risks and propose mitigations.

Risks of widely available model weights

This section addresses the NTIA RFC Question 2(a): What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?

We identify two classes of risks: direct risks, where an individual AI model poses some harm to society, and indirect risks, where an AI doesn't directly cause harm to society, but does lead to harm down the road.

Direct Risks

Weaponization. AI technology could be weaponized to cause massive harm. If an AI system has the capability of a nuclear engineer and is widely avialible, this could dramatically accelerate the development of nuclear weapons. Even narrow AI systems may provide substantial weaponization concerns, particularly in CBRN domains. There are substantial concerns both in designing more dangerous weapons (e.g. developing a supervirus that is substantially deadlier than naturally occurring viruses) and in reducing the barrier to entry (e.g. reducing the number of scientists to build a WMD from a team of 50 to a team of 5).

Loss of Control. If we develop sufficiently capable autonomous AI systems, we may face widespread loss control of these AIs. This is a particular concern when developing extremely intelligent AI systems that exceed human capabilities across a wide range of tasks. **If AI**

systems which can autonomously do AI capabilities research become widely available, it will become extremely difficult to prevent the development of substantially smarter-than-human AI systems. Widely releasing these AI model weights give a large number of actors access to these models, and each actor will have a large financial incentive to build more capable AI systems. Even if there are substantial safety concerns, it is unlikely that all of these actors will be able to coordinate to avoid building the extremely capable systems, which are also the most dangerous.

Open sourcing AI models that pose severe direct risks such as weaponization or loss of control is a mistake for several reasons:

1. **We allow widespread access to highly dangerous technology.** For both AGI systems and WMD-like AI, giving widespread access dramatically increases the likelihood that someone will use them irresponsibly. Particularly concerning applications include the development of extremely competent agentic AGI systems, AI-engineered pandemics, and large scale AI powered deception.
2. **AI safety safeguards are rendered obsolete.** It's extremely easy to remove safety features from an open-weights AI model. We can expect any safety techniques that are applied to be removed by people wishing to do unsafe things with their AI system that are limited by the safety techniques.
3. **We lose the ability to turn the AI off**. If an AI is released throughout society, there's no clear way to turn it off: you can't get people to un-download the piece of software. In an emergency, an individual company can take down a dangerous product, but a decentralized AI would be impossible to shut down. This strongly limits our ability to respond to threats.

Indirect Risks

AI capabilities acceleration. Many experts are concerned that building and empowering highly capable AI systems could lead to very bad outcomes, including extinction. We agree that society is not sufficiently prepared to build highly capable AI systems. We lack both technical safety solutions for aligning an AI system to its developer's interests and government response capacity to react to the drastic changes in the rate of technological development such as widespread job loss and problems associated with the obsolescence of human workers. Open weight models can accelerate the development of extremely powerful models, giving society less time to prepare.

Reducing the American lead. The duration of time between when the first actor who can build AGI and when adversaries can build AGI is critical, as this is a period when urgent technical and governance measures must be taken to improve the safety of AI systems. Releasing models that push the open-weight frontier reduces this gap.

Policy Recommendations

Our two main policy recommendations around open-weight AI systems are:

1. **Improve monitoring and forecasting capacity** to understand the timeline until AI systems pose direct risk.
2. **Secure future high-risk AI models** by creating capability based safety standards that involve preventing the model weights of high risk AI models from becoming widely available.

Improve AI monitoring and forecasting capacity

AI development is highly unpredictable, and so rapid policy responses may be necessary. Monitoring and forecasting AI development is critical for making informed policy decisions. Ideally, we would develop precise measurements that could evaluate the extent to which individual models pose risks.

Model Evaluations are tests which attempt to measure certain properties about an AI model.

The simplest form of model evaluation involves simple benchmarking to check performance on relevant datasets. For example, benchmarks for highly capable models include:

1. GPQA (A Graduate Level Google Proof Q&A Benchmark) is a dataset containing expert level multiple choice questions in chemistry, biology and physics.
2. SWE Bench contains unresolved software issues from GitHub, and checks whether AI models can resolve them.
3. GAIA is a benchmark containing tasks that involve tool use and many step reasoning. This tests models capability to solve complicated problems.
4. Weapons of Mass Destruction Proxy (WMDP) is a benchmark that tests AI models' ability to answer questions relating to the development of WMDs, including cybersecurity, bioweapons, and chemical weapons.

One can also conduct more sophisticated evaluations that test models in more realistic settings. For example, METR (formerly ARC Evals) made an evaluation that tested AI's capability to autonomously set up new instances of itself on new servers. While model evaluations are fallible, they provide useful information about the level of risk from AI models that are being developed.

Model evaluations attempt to address direct risk from specific AI systems. However, it is difficult for model evaluations to capture structural effects like speeding up AI capabilities or reducing the lead time of the leading actor. In order to estimate these effects, it is important to **forecast the danger** coming from future models. If there is a significant chance that the next generation of AI systems will be high-risk, we must take steps to contain them before they are widely released. To do this successfully we need accurate forecasts of future model capabilities. This forecasting should involve consulting with professional risk assessors and AI experts in order to estimate critical strategic variables.

Capability Based Safety Standards

To address risks, the government could require that model developers follow safety procedures to evaluate and counteract risks. These could either be written into voluntary standards such as the NIST Risk Management Framework, or ideally, used within required safety standards enforced by an AI regulator. Before releasing models, these safety standards could specify which types of evaluations to run for estimating certain types of harms.

One part of these safety standards could involve evaluating the effects of giving widespread access to model weights. In this section, we sketch how this evaluation could work. Because the risks of AI models depend on the capabilities of the underlying system, we propose three tiers that could be used in safety standards: low risk, medium risk, and high risk. The development of complete standards should involve setting specific thresholds using the AI capability evaluations discussed above along with setting up safety and security requirements for AI models at each capability level.

Low Risk	**Medium Risk**	**High Risk**
AI systems that are less capable than the most capable open weight models Narrow AI systems, e.g. recommender systems	AI systems that accelerate the development of high-risk AI systems AI systems that could be enhanced to become high-risk	Artificial General Intelligence (AGI) WMD-like AI

Low Risk. Given that models with higher capabilities are already widely accessible, the risks of widely releasing models below the current open-weight frontier are very low. Given that there are benefits and few risks, AI safety standards should allow model weights to be released freely.

Medium Risk. Releasing model weights that are more capable than existing open-weight models pose some risk, even if they appear safe during the evaluations testing for specific dangerous capabilities. Medium risk models pose:

- **Direct risk** because the tests are fallible: a model that appears safe during initial testing might turn out to pose dangers later (after it is too late to recall the model) based on post-training enhancements, a wider availability of plug-ins, and a decreased cost of compute.
- **Indirect risk** because any model that pushes the capability frontier will have effects on other AI development, including reducing the American lead time on AI because of accelerating other AI projects.

Current AI models on the open weight frontier are probably not sufficiently capable that they pose significant risk. We therefore recommend thorough monitoring of these systems for risks that may emerge, but do not yet recommend enforcement actions taken against models in this category.

High Risk. High risk models are models that pose substantial risk because they have dangerous capabilities or are soon projected to have highly dangerous capabilities due to post training enhancements. The threshold for being a high risk model should include tests for both weaponization capabilities and loss of control capabilities. The model weights for high risk models should not be widely released because of the substantial risks described above. Fortunately, no current AI systems are in this high risk category.

Selected Questions from the NTIA Request for Comment.

Is it possible to generally estimate the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks? How do you expect that timeframe to change? Based on what variables? How do you expect those variables to change in the coming months and years?

We find that the timeframe between closed and open models right now is around 1.5 years. We can arrive at this conclusion by analyzing benchmark performance between current leading open weight AI models and the best closed source AI models.

One of the best open weight AI models is Meta's Llama-2 70B, which has comparable benchmark performance to GPT-3.5.[2] Llama 2 finished training in July 2023[3], right before it was released. GPT 3.5 was released after a period of red teaming on March 15 2022, and so likely finished training at least several months before, likely Jan 1 2022. This gives us a performance gap of 1.5 years.

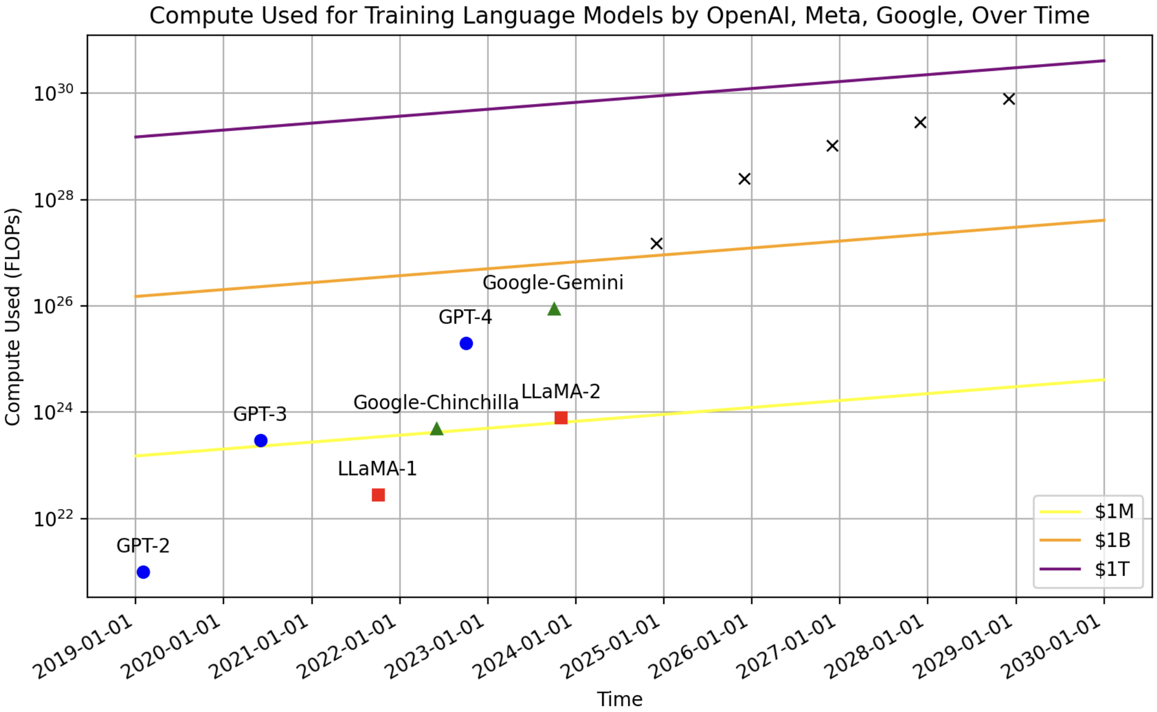
There are multiple competing effects that determine how this timeframe will change over time. Here, we survey some of the major considerations: the cost of training runs, ideological commitments, feedback loops, and the proliferation of information.

The cost of training runs. Large dollar costs will incentivize LLM developers to avoid widely releasing their model weights. GPT-4, trained in 2022, cost over 100M. The training costs for the largest AI model has been increasing at a rate of 3.1x/year. If we extrapolate this trend, we get very large dollar costs for frontier AI models over the next few years. Quantitatively, we can graph the compute used for existing models over time, and then extrapolate the growth using the trendlines. However, naive extrapolation is probably incorrect because the rate of growth is unsustainable: AI developers will run out of capital to spend on training runs. Large tech companies will probably be unable to raise $1T to spend on individual training runs – they do

[2] Llama-2 has a performance of 1082 on LLM-Leaderboard and 30.4 on GPQA (extended set), which is slightly better than GPT-3.5, which has 1068 on LLM-Leaderboard and 28.2 on GPQA (extended set).

[3] See https://github.com/microsoft/Llama-2-Onnx/blob/main/MODEL-CARD-META-LLAMA-2.md

not have that much cash on hand[4] and there are few funding sources which could be that large. Additionally, at that point they will run out of hardware to buy.[5]



This chart shows some existing large language models, as well as lines showing the amount of compute[6] that can be bought for $1 million, $1 billion, and $1 trillion, adjusted as compute gets cheaper over time. The unlabeled x's represent our all-things-considered best estimate for the compute used in the largest closed training run over the next several years. Over the next two years, this is a linear extrapolation from the past, and after that we adjust downward to account for running out of hardware.

Some AI developers have strong commitments to open source. Mark Zuckerberg laid out a vision of Meta open sourcing the models all the way to AGI. This means that some AI developers may widely release AI model weights despite financial cost or substantial risk to public safety. This underscores the need for regulation: model developers may not take appropriate safeguards on their own.

Feedback loops from closed source AI R&D could extend lead times.

[4] Google had 110B cash on hand in December 2023.

[5] A rough calculation is that Nvidia produced 1.5e6 H100 GPUs during 2023, each of which produced 1e15 FLOP/s in FP16, which together make up 1/10th of the global compute supply. Assuming that GPU utilization is 50% the entire compute supply could support a training run of 6e28 FLOP. Given that GPT-4 used 2e25 FLOP, it used about 3e-4, or .03% of the world's compute supply. This limits the amount that increased investment could result in increased compute without increasing global chip production to around 1000x without additional chip capacity. Chip supply is highly inelastic given the 3-5 year lead time to construct a new fabrication plant.

[6] Compute is measured in floating-point operations (FLOPs), assuming that training is done using a bit-length of 16. If training is done in a different bitlength, one must adjust the FLOP availability proportionally, e.g. for 8-bit training, there are twice as many FLOPs available.

AI models can accelerate AI research – there are many existing examples for AI models providing value to accelerate research. Model developers that keep their models closed may be able to extend their lead by accelerating AI research. This will become a larger and larger effect over time as AI models become increasingly capable. When AI systems are able to provide useful research contributions without humans in the loop, there will be a substantial increase in the rate of research progress because of the large increase in the number of researchers able to tackle specific problems that arise.

Diffusion of algorithmic insights. AI labs historically have had weak information security – there have been a plethora of leaks of algorithms, as well as plenty of personnel flow between labs. In the runup to AGI, poor information security will probably prevent the OS frontier from falling too far behind the closed source labs.

What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights, or limit their end use?

To reliably prevent wide availability, it is necessary to (a) not widely release model weights and (b) implement good information security practices to prevent model weights from being exfiltrated by adversarial actors.

If the model is widely available, it is difficult to limit the end use of models. On specific tasks that require many inference steps (e.g. if the user is trying to automate certain jobs on an ongoing basis), limiting those tasks is possible by restricting hardware usage. However, most usages of models can be done with very small amounts of hardware. To give an idea of the scope, to use GPT-4, OpenAI charges 10$ per 1M tokens. The total amount of computation in a single forward pass is simply quite cheap to create, and therefore does not require all that much compute.

As an alternative to releasing model weights publically, there are partially open measures that obtain some of the benefits of open weight models while mitigating some risk. For example, one could use the NAIRR as a secure compute cluster to give access to some specific users who are doing research on an AI model with dangerous capabilities without making the weights widely available to arbitrary actors. In particular, this can allow us to continue accelerating AI safety research and positive AI use cases, while preventing malicious or reckless actors from obtaining the model weights.

Noting that E.O. 14110 grants the Secretary of Commerce the capacity to adapt the threshold, is the amount of computational resources required to build a model, such as the cutoff of 10^{26} integer or floating-point operations used in the Executive order, a useful metric for thresholds to mitigate risk in the long-term, particularly for risks associated with wide availability of model weights?

Floating point operation (FLOP) thresholds are useful thresholds in the toolbox, but are limited in some important ways. FLOP thresholds are clear and easy to measure. Moreover, they are a reasonable proxy for capabilities: models that increase in scale tend to also increase in capabilities.

However, AI algorithms are improving every year. With a given amount of compute, developers will be able to build more and more capable models over time. Additionally, the physical bounds on AI capabilities suggests that it is physically possible to build AI systems that have advanced capabilities using significantly less than 10^26 FLOPs.[7]

Another option is to use a capability-based threshold instead of a compute based threshold. The key advantage of this approach is that the capabilities of an AI system are the true source of risks, and so this can be a much better proxy.

Unfortunately evaluating model capabilities has proved challenging, and so this is less clear as a regulatory threshold than FLOP. AI capabilities are rapidly increasing, and so benchmarks are quickly becoming saturated. A capability based evaluation will need to be updated often.

[7] As a rough calculation, a human brain uses around 10^15 FLOP per second, though various methods find different numbers. 10^9 seconds is around 30 years, and so the total amount of computation done by a human brain running for 30 years is 10^24 FLOP.