Machine Intelligence Research Institute
Berkeley CA

March 27, 2024

National Telecommunications and Information Administration
U. S. Department of Commerce
Washington DC

Re: NTIA AI Open Model Weights RFC (NTIA–2023–0009)

Our organization, the Machine Intelligence Research Institute (MIRI), is focussed on increasing the probability that humanity can **safely navigate the transition to a world with smarter-than-human AI**. We investigate ways to mitigate the **catastrophic and societal-scale risks** associated with artificial intelligence (AI) systems as they near and surpass the capabilities of humans.

As such, our answers below will mainly relate to the risks posed by powerful autonomous AI systems, in particular the risk of losing human control over such systems, as well as the potential risks arising from the development and use of powerful models by incautious or malicious actors. There are many other important risks posed by AI systems, however we generally won't speak to those risks, as this is not our area of expertise.

We believe that current AI systems likely do not pose such risks, and this is primarily due to their lack of capabilities, i.e, the systems are not powerful enough to cause catastrophic harm. However, many experts believe that future AI systems will become much more capable, removing the justification for considering them safe.

In this Request for Comment (RFC) response, we primarily want to help in setting a good precedent for future systems, where it is only appropriate to release model weights of AI models if one can be confident that this will not cause significant harm.

Our main policy suggestions are related to understanding and mitigating risks before AI systems are deployed or released. We believe that current understanding of the inner workings of AI systems and the best currently available risk mitigation strategies are strongly insufficient relative to the magnitude of the risks posed by future powerful AI systems.

We would like to highlight a few technical and strategic considerations relevant to the

question of making model weights widely available.

There are currently no methods for reliably determining the capabilities of AI systems before they are trained—or even for ruling out capabilities after training.

As the CEO of OpenAI Sam Altman recently said, before a model is trained, predicting its capabilities is a "fun guessing game." Additionally, post-training enhancements such as novel prompting techniques, scaffolding, and fine-tuning have been known to increase the range of capabilities displayed by models in ways that are difficult to predict in advance. This means that if a model's weights are made widely available, wider experimentation with the model may unlock unexpected capabilities. Even if a model was initially considered harmless, such enhancements may increase its capability profile beyond acceptable risk thresholds.

A final point to stress is that the release of model weights is an irreversible action; once model weights are released, any actor can use them, no matter how malicious or incautious.

Below we respond to specific questions from the RFC.

1a) Is there evidence or historical examples suggesting that weights of models similar to currently-closed AI systems will, or will not, likely become widely available? If so, what are they?

There is significant evidence that model weights of powerful AI systems will become widely available, especially in the absence of regulation. The primary evidence for this is AI developers releasing the weights of such models multiple times since the development of large language models (LLMs). One of the first LLMs, GPT-2, had its weights released by OpenAI (which would eventually go on to create ChatGPT), and this trend has continued into the present day with xAI (Elon Musk's AI company) releasing the model weights to their LLM Grok-1. In addition to these, there are many other examples. One of these examples, Mixtral 8x7B, appears to be competitive in capability with GPT-3.5. Given that GPT-3.5 was released in March of 2022 and Mixtral 8x7B was released in December of 2023, this (at least naively) suggests that open foundation models may lag behind frontier models by approximately two years, and that weights of models similar to currently-closed AI systems may be available by early 2026.

There is also precedent for model weights "leaking", i.e., being made widely available

without the developer's consent. Prominent examples include the leak of Meta's LLaMA-1 model weights, and recently AI developer Mistral reported a leak with a link to their model weights being posted anonymously to 4chan.

Furthermore, as machine learning hardware continues to improve in power and price-efficiency, and as algorithms to train language models continue to advance, requiring less computation to reach the same level of performance, we should expect more and more actors to be able to train their own models of comparable capabilities to currently-closed frontier AI systems. This will directly enable such weights to become widely available by enabling a wide variety of actors to train and share such models. This includes actors that may share model weights as part of a business strategy to ensure their models are widely integrated into other tools, or due to a belief that it is important for the weights of widely-used models to be widely available.

1d) Do certain forms of access to an open foundation model (web applications, Application Programming Interfaces (API), local hosting, edge deployment) provide more or less benefit or more or less risk than others? Are these risks dependent on other details of the system or application enabling access?

For current and near future models, risks can be reduced by monitoring their usage. This includes monitoring of a user's queries to an AI system, the AI's outputs, and any fine-tuning of the system. This monitoring can potentially be done automatically by other AI systems to preserve a user's privacy. Such monitoring is not feasible if the model weights are widely available, because users may simply be able to host the models locally without any monitoring. Additionally, it has been demonstrated that with access to model weights "safety fine-tuning" (such as training a model to refuse harmful requests) can easily be removed with minimal cost. **Therefore, once model weights are widely available, different forms of access provided to the model have little bearing on risk.** (We elaborate on this below in our answer to question *5d*.)

Even where monitoring is possible, there remain risks from sophisticated attacks by users or hostile actors including via web application or API access. This could be even as simple as "system jailbreaks", where targeted prompting by a user can get the system to reveal hidden content. This becomes less likely the stronger the security monitoring is. However, it is difficult to be confident that security monitoring is exhaustive, since it may well be easier for attackers to find ways to jailbreak models than for model owners to prevent jailbreaks.

Weights will tend to be more secure when they are stored in a smaller number of places, each of which can have more security measures taken. If models are deployed more broadly, such as on consumer devices, it becomes more likely that their weights will be widely released. This is because a broader deployment of weights increases the number of actors with licit access who would be able to share the weights, as well as increasing the attack surface for agents attempting to illicitly access the weights. Deploying on consumer devices may also allow users to evade monitoring by, e.g., using the model offline.

2a) 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?

For current AI systems, we believe that the main risk of catastrophic or societal-scale harm comes from malicious actors using these AI systems to cause harm, e.g., with CBRN (chemical, biological, radiological and nuclear) weapons, cyber attacks, misinformation or persuasion campaigns, etc. These AI systems may be more useful to malicious actors than other resources such as search engines, for instance by customizing the presentation of information to be more comprehensible to specific users, and combining information in ways a search engine cannot. They could also automate the execution of cyber attacks and misinformation campaigns. By reducing required human involvement in the information gathering or execution stages of these attacks, AI systems could significantly decrease their cost.

Additionally, one effect of making training data and source code more available could be to accelerate the rate of AI progress by incautious or malicious actors by giving them more powerful models to build off of and enabling them to use more efficient architectures and training methods. For example, this could lead to foreign militaries and oppressive regimes gaining the ability to create more capable AI systems, which they could then use to cause harm in the ways described above.

Future AI models (including potentially in the near future), may only be able to be safely used with significant security precautions and safety practices. These may be necessary to prevent "self-exfiltration" where an autonomous AI system copies its weights to a location where it can no longer be safely monitored and secured—one of the primary risks tracked by major AI developers such as OpenAI, Anthropic and Google DeepMind,

as well as METR, a leading AI evaluations organization. **Models capable of self-exfiltration likely pose greater catastrophic risks from loss of human control.** Major AI developers *may* be able to eventually develop and implement the appropriate security and safety practices to prevent self-exfiltration, but **if weights of a powerful AI model were made widely available, smaller or less security-conscious actors would be unable or less likely to take adequate precautions.** It is important to note that this risk primarily applies to AI systems which are capable of self-exfiltration, which likely do not include current models.

2f) Which are the most severe, and which the most likely risks described in answering the questions above? How do these set of risks relate to each other, if at all?

The most severe near term effects are likely enabling CBRN threats (e.g., a malicious actor using AI systems to design an engineered pandemic) and cyber attacks on critical infrastructure.

For future AI systems, there is a concern shared by a significant portion of the research field, including industry and academic leaders, that advances pose a risk of loss of human control, potentially leading to catastrophic outcomes such as human extinction.

The related risk of self-exfiltration may require an AI to have strong capabilities in coding and finding cybersecurity vulnerabilities, or in human manipulation and persuasion. These abilities are directly related to ways in which incautious or malicious actors could cause harm.

Finally, we do want to acknowledge that there are risks associated with limiting access to model weights and training code. Open access to these systems allows a much larger set of external researchers to work with foundation models to better understand their capabilities and inner workings, and conduct and contribute to essential research related to mitigating these risks. In our view, this is an important consideration in favor of current models being made more open, and investing in 3rd alternatives that could allow these researchers to continue doing this essential work without direct access to model weights of future more capable models (which we elaborate on in our answer to question *8a* below).

5a) What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?

We will consider the cases of misuse by malicious actors and harm caused by autonomous AI systems separately, as these will likely require different model evaluations.

For misuse risks (i.e., malicious actors using an AI to cause harm), one can gain confidence that model weights are safe to release if they are demonstrated to not meaningfully assist humans with dangerous tasks (such as developing biological weapons). Such evaluations may involve having a human evaluator attempt to use the model (including all feasible post-training enhancements) to assist with such tasks. If the evaluator is unable to make the AI system meaningfully assist them, and the evaluator has good reason to believe that they are eliciting the model's full capabilities, then the model may be considered safe. It is important to note that this likely requires testing the model for a range of dangerous behaviors, and also that there may be undiscovered techniques that would make the model more capable that have not been discovered or widely publicized yet. This should include testing done by external evaluators. Models that are able to meaningfully assist malicious actors to (or enable incautious actors to inadvertently) cause significant harm should not have their weights be made widely available.

Different evaluations are required to gain confidence that an AI system would not be able to autonomously cause significant harm. As models become more powerful, they may gain more "situational awareness", and be able to realize they are being evaluated and behave differently. A striking early example of this behavior has been observed in the recently released Claude 3 Opus model from Anthropic. If models are able to detect when they are being evaluated, they may behave differently in their evaluations and in deployment. This means that the evaluations may no longer be an accurate representation of the model's true capabilities and behavior.

More research is needed to develop principled evaluations of AI systems' true capabilities. This research could include attempting to mechanistically explain how models are able to perform tasks, and then gaining confidence that models don't have the mechanisms required for certain dangerous behaviors (e.g., a model could be lacking the ability to model human behavior, or to detect cyber security vulnerabilities). This may become much more difficult if the model weights were made widely available, as malicious or incautious actors may be able to fine-tune the model (at a fraction of the cost it took to train them) to give them these dangerous capabilities. Even if we are able to gain confidence that a model lacks certain capabilities (which is not currently feasible), the ability to fine-tune the model would likely remove these guarantees.

There is not currently a rigorous way to evaluate the costs and benefits of making model weights widely available. This evaluation requires understanding the dangerous capabilities of models, and there is currently no way to know the true limits of these capabilities, especially when we consider that novel post-training enhancements will likely be developed in the future.

Current model evaluations only cover an extremely small fraction of what models may be able to do, especially when acting autonomously. We also believe that the threat modeling and risk assessment for this area is still nascent, and so we do not have confidence that the range of risks caused by the release of model weights has been properly considered.

5d) Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?

There is some research on *hardware-specific models* and *self-destructing models*. Hardware-specific models are AI models which would only run on certain restricted hardware (e.g., one could imagine OpenAI models only running on OpenAI-specific GPUs). This means that even if the weights of a hardware-specific model leaked, the model would still not be usable as it would also require the specific hardware to run.

Self-destructing models are models which are trained in such a way that they are particularly difficult to fine-tune on specified tasks. This is a promising area of research. Potential future developments which obviated the need to manually specify dangerous tasks to prevent fine-tuning on could plausibly make models that initially lacked dangerous capabilities safe to release, although they would not enable the release of models that natively possessed dangerous capabilities.

If methods such as these are not employed, then there is likely no way to "regain control over" or "restrict access to" models which have widely available weights. If model weights are released widely and are able to be used and fine-tuned, it will likely be extremely difficult to control what they are used for or to restrict access.

7i) Are there effective mechanisms or procedures that can be used by the government or companies to make decisions regarding an appropriate degree of

availability of model weights in a dual-use foundation model or the dual-use foundation model ecosystem? Are there methods for making effective decisions about open AI deployment that balance both benefits and risks? This may include responsible capability scaling policies, preparedness frameworks, et cetera.

Our answer to this question will discuss limitations of frameworks such as Anthropic's Responsible Scaling Policy (RSP) and OpenAI's Preparedness Framework (PF), as well as the sorts of questions which would need to be addressed in order for these frameworks to allow one to be confident in the safety of an AI model.

The RSP and PF attempt to rank models according to how capable and dangerous they are, e.g., in the RSP current models are ASL-2 (AI Safety Level 2), and near-future models will likely be ASL-3. These frameworks attempt to achieve safety by implementing appropriate security and safety measures according to how capable the models are. This will involve extensive security and monitoring, and safety fine-tuning to prevent malicious actors from using the models to cause harm. However, without these safety and security measures future models would very likely be dangerous and able to cause significant harm. If the model weights were made widely available there would likely not be any security measures and any safety fine-tuning could be easily removed. Following the RSP and PF frameworks, for sufficiently powerful models it will not be appropriate to make the weights widely available.

The RSP and PF currently rely on evaluating a model's capabilities, and this means there is currently a lot of guesswork in these frameworks. In order to be confident that an AI system is safe (even with strong internal security measures), we believe that AI developers must be able to adequately address the following questions with scientific rigor:

- Are the evaluations measuring the model's full capabilities? (this is sometimes referred to as capabilities elicitation)
- Do the proxy tasks which comprise the evaluations actually measure the dangerous capabilities of interest?
- Do the dangerous capabilities which are being evaluated exhaustively cover the ways in which the model could cause unacceptable harm? How might this change in the future?

An important part of these frameworks is attempting to predict the capabilities of future models, in order to avoid training models for which there are insufficient security measures.

To have confidence that the models being trained are not going to "overshoot" the security measures, it is important to have a principled way to predict model capabilities. There is currently no way to do this, and model capabilities are known to emerge unexpectedly with scale. It is important to ensure that before training powerful models, we understand their likely capabilities well enough to implement security measures that prevent the models from posing catastrophic risks.

8a) How should these potentially competing interests of innovation, competition, and security be addressed or balanced?

We believe that AI developers should not be allowed to impose extreme risks onto the non-consenting public.

Future advanced AIs have the potential to pose catastrophic risks which harm large numbers of people, including causing human extinction. It is difficult to be sure of the capabilities possessed by any given model, especially prior to training it, and currently known measures to reduce these risks have serious flaws. We believe that it is critical for the government to protect its citizens, to develop clarity about the risks which these AI systems pose, and to prevent AI systems from being released when not in the public's best interest.

Race dynamics and other competitive pressures between AI developers may cause them to not prioritize the safety and security of the AI systems which they develop. **It is important that governmental or other regulatory authorities are able to monitor and audit AI developers to ensure they are not lax on safety and security.** There can be important societal benefits from AI, but these should not come at the cost of safety from catastrophic harms, and especially **should not expose the public to large-scale risks they did not consent to.**

One way to promote safe research into powerful AI models may be for the government to invest in and promote forms of limited access to the internals of powerful AI models to researchers (e.g, through the National AI Research Resource). This would enable these models to be studied, promoting research on their capabilities and how they can be made safer without allowing the model weights to proliferate in an uncontrolled way. The United States could develop better frameworks for these types of access, incentivize developers to grant these forms of access in ways that minimize risk of full model weights leaking while allowing safety research and assessments to take place, and thereby help democratize this form of science.

8b) 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 1026 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?

Metrics based on the amount of computation used to train a model (such as the number of 10^{26} integer or floating-point operations) currently appear to be adequate for the purpose of determining whether models are advanced enough for it to be valuable to share information pertinent to their safety.

However, these metrics are likely to become inadequate in the future. **Due to progress in machine learning algorithms, the amount of computational resources required to train an equivalently powerful model halves approximately every 8 months. This means that thresholds based on the resources used to train a model will need to be brought down over time**, to prevent dangerously capable models slipping under the threshold.

Furthermore, it may be difficult to adjust these regulatory thresholds fast enough, due to the rate of technological progress outpacing the speed of the regulatory reaction. This may result in the regulations always trying to catch up with the fast-moving technology. To partially alleviate this, **we propose that the relevant regulatory body should have the ability and the mandate to quickly modify these thresholds in response to new developments, and also to base these thresholds on the forecasted future state of the technology** rather than only considering the state of technology in the present.

Potentially more importantly, these thresholds based on these metrics do not account for sudden breakthroughs in AI capabilities, and so would fail if a novel discovery let an AI developer train an equivalently powerful model with far fewer resources. A preferable metric would involve measuring and predicting, in a principled way, an AI's ability to do dangerous tasks and become powerful.

The lack of such a preferable metric should not be an excuse to continue to ignore the limitations of the metric of training computation. Instead it should motivate research into developing metrics which would allow us to develop powerful AI systems in a safe and principled way.