Before I begin, I'd like to thank you for accepting this rather lengthy comment, and I apologize in advance for posting it anonymously. To be fair, there was bold text on the submission page specifically stating "***Do not submit personally identifiable information through this form***", so it seemed smart to heed that warning.

With that said, I'll do my best to introduce myself all the same.

I am a software development manager working in Fintech with a master's degree in computer science and 12 years of professional experience in my field. Last summer I began learning about Open-Source AI as a hobby and have been both intrigued and emboldened by the technological innovations I have seen within that space. I have spent the past year tinkering, learning and generally trying to better understand not only how Open-Source AI works, but also how it is perceived and utilized within the tech sector in comparison to its proprietary closed-weight brethren.

I am not, by any means, a Machine Learning expert nor an AI scientist. My knowledge of AI is purely that of a hobbyist. However, I do feel that my tech background and focused interest in this field may allow me to make useful and helpful contributions to this discussion. I know that my response is lengthy, but I do hope that you'll take the time to read over it.

I would like to note that my answers in this document are mostly relegated in topic to Text Generation AI. I do not feel that I have an acceptable level of knowledge or understanding on the technology and social topics related to image generation AI to properly speak to those. Some answers may overlap, but please read each answer with that in mind.

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

The terms "Open" and "Widely Available" require two separate definitions. For example, any software released under the MIT License on a hosting platform like Github could be considered "Open", as the license allows free use to anyone for both commercial and personal needs (https://pitt.libguides.com/openlicensing/MIT). In comparison, the Llama license for models hosted on HuggingFace.co also offers such usage. So in terms of a model being "Open", one could clearly define open as being any model that has Licensing which allows free use with limited enough restrictions that it can feasibly be applied to personal and/or commercial purposes.

Additionally, there is a further layer to the "open"ness of a model. One can also consider the release of the "weights" when determining whether a model is open. Without the

weights, a model cannot be trained to add new information, nor can many (if any) methods be employed to determine how the model comes to the answers that it does. Can one truly consider an immutable black box piece of software to be "open"? Probably not.

With that said, "Widely Available" should be defined separately from "Open". Widely available should consider not just availability to download, but availability to utilize. For example: Falcon 180b would be considered both "Open" and "Widely Available" without this consideration, while in reality the vast majority of people in the United States cannot run this model.

To help with quick napkin math, consider an "unquantized" or "raw" version of a model, an FP16 model, to require 2 gigabytes (GB) of Video RAM (VRAM) per 1 billion parameters to utilize. The raw models are the highest quality that you can get with a model. This means that Falcon 180b could require over 360 GB of VRAM.

For comparison's sake, one of the most expensive and top-end modern graphics cards available to consumers, the RTX 4090, has 24GB of VRAM at a cost of almost $2,000. The average American does NOT have an RTX 4090 available to them. According to Valve, the owner of the platform Steam which is a popular computer gaming platform with over 120 million monthly active users as of 2021, the most popular graphics card is the RTX 3060 (https://store.steampowered.com/hwsurvey/videocard/). The RTX 3060 has 12GB of VRAM. Going back to the Falcon 180b, which raw requires over 360 GB to run, we can see that the average user doesn't come close to being able to run this model.

According to the ESA (https://www.theesa.com/news/video-games-remain-americas-favorite-pastime-with-more-than-212-million-americans-playing-regularly/), 65% of the US population are video gamers. Being very liberal with estimates, we will falsely assume that ALL of them are computer gamers, as this would reflect poorly against my own argument and thus likely be more fair in terms of me making assumptions.

So, assuming that 65% of all US citizens had a gaming computer available to them, we can then see that 10% of those gamers (again, at best) have an RTX 3060. This means that some 6.5% of the country has at most 12GB of VRAM available to them.

And those "big" cards with 24GB? The expensive ones? 0.90% of gamers on Steam have the RTX 4090, or 24GB of VRAM, to work with. That's what... 0.6% of the US population? Less?

There are, of course, alternatives. An Apple Mac Studio is capable of running much larger models, with 147GB of VRAM naturally available on the $6,000-$8,000 M2 Ultra Mac Studio with 192GB of RAM. But again, less than 1% of all Macs sold were Mac Studio models (https://9to5mac.com/2023/01/09/apples-most-popular-mac/). And Macs only account for 25% of all desktops in the US (https://gs.statcounter.com/os-market-share/desktop/united-

states-of-america). This means that 1% of 25% of US desktops MIGHT be capable of running relatively large models, since not all Mac Studios come with 192GB of RAM. In fact, many likely have far less.

Another alternative is GPU rentals, but costs can be quite prohibitive there, and they are challenging to work with. Consider RunPod, one of the most popular and affordable solutions in the space: at the time of this writing, it costs almost $4 an hour for 80GBs of VRAM (https://www.runpod.io/gpu-instance/pricing). At this price, using that tier for just 3 hours a day would cost an upwards of $350 a month to use. And it still could not run an unquantized Falcon 180b.

Now, software solutions have also become available. These raw models can be "compressed", or quantized, into much smaller and more usable sizes. The largest quantization available is ~8 bits per weight (bpw), which comes out to about 1GB for every 1b of a model. So at 8bpw, the Falcon 180b would be around 180GB. Still too large for an 80GB card.

We could further compress the model down to a much more modest 2.55bpw, one of the smallest available quantizations, which would reduce the model down to ~58GB ((2.55/8)*180). Including overhead for things like KV Cache (which would take up around 3GB or so), you COULD fit that on your $4 an hour rental; however, when a model is compressed that heavily, the quality is reduced so much that the model becomes almost unusable, and its responses can often be confused and factually inaccurate.

So, with all of this said: the question, then, is whether you can consider a model "widely available" if *at best* less than 1% of the entire US population can use it. I would argue to say "no".

> **a. 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?**

First, it is important to understand that the blanket of "currently-closed" AI is a bit over-broad. Consider the recently closed-sourced model Grok, produced by xAI (Twitter/Elon Musk). That model is a ~314b model (https://arstechnica.com/information-technology/2024/03/elon-musks-xai-releases-grok-source-and-weights-taunting-openai/). Grok used to be closed-source, and has now been opened as of the week of March 8th. By your definition, it WAS a currently-closed model that is now Open.

However, Grok is not to be conflated with ChatGPT 4 level models for two reasons.

- First: ChatGPT 4 is rumored to be a 1.7T, or 1.7 Trillion, parameter model (https://twitter.com/swyx/status/1671272883379908608). That puts it being ~5 times larger than Grok. And yet both would be categorized as the same "currently-closed" type of model before Grok was released.

- Second: By most leaderboards, Grok does not come close to the capability of Claude 3 or ChatGPT-4, ranking much closer to currently open-source models in the 50-70b range for most topics (https://www.vellum.ai/llm-leaderboard).

Further, let us consider the previous discussion on "Widely Available". If we determined than less than 1%, at best, of US citizens can use a HIGHLY compressed and very poor-quality version of Falcon 180b, how many people do we expect to be able to use a 314b model?

In reality, there is currently no historical example or evidence of a "currently-closed" model like ChatGPT-4 becoming Widely Available. Any model with the capability and size of ChatGPT-4 are not only NOT open, but even if they were released tomorrow, they would still be off the table for the vast majority of people. You could likely count the number of non-commercial entities capable of running even a quantized version of ChatGPT 4 on your hands.

b. 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?

The speed at which technology is progressing in this industry makes this question challenging to answer. The rapid progression in the past 2 years causes any historical example to go out the window. Allow me to explain:

Consider ChatGPT 3.5: it was released to the general public on April 27, 2023, and is a Currently-Closed model. Llama 2 is an Open Weight and Widely Available family of models, the largest of which competes directly with ChatGPT 3.5 (https://www.vellum.ai/llm-leaderboard). Llama 2 was announced on July 18, 2023, a mere 3 months later.

However, looking at this same leaderboard, we can see that neither model comes close to ChatGPT-4 capability. ChatGPT-4 released March 14, 2023, and has yet to see an open weight equivalent in terms of size or quality. So this brings us back to the definition of

"currently-closed" being overbroad. The "letter of the law" response to your question is 3 months; but the real and intellectually honest answer is that there is no historical evidence with which to base an answer to this question, because it has not yet happened.

c. Should “wide availability” of model weights be defined by level of distribution? If so, at what level of distribution (e.g., 10,000 entities; 1 million entities; open publication; etc.) should model weights be presumed to be “widely available”? If not, how should NTIA define “wide availability?”

Wide Availability should only, and can only, be quantified in terms of how many entities can legitimately use the thing once it has been made open.

First – consider that once something is released to the internet, the number of entities that can currently be holding a copy of that thing is N, where N is less than or equal to every human being with internet access and the available hard drive space across the entire planet.

You also cannot rely on download counts from a website, as once something is downloaded, it can be uploaded to other places, shared via Peer to Peer (p2p) programs like Torrents, and otherwise distributed through mostly untraceable means.

Second – what’s the value in a file you can't open? Most computers in the US would have the hard disk space to store a copy of Grok-1, the 314b model. Far less than 1% of those computers can run that model. What's the point of calculating the number of people with a digital paperweight taking up space on their hard disk, as opposed to calculating the number of individuals who can actually make use of it?

I would consider "Widely Available" as fitting within the specs to be accessed by a large number of individuals. For myself, personally, I'd consider 7b models to be "widely available"; anecdotally you can see this in online communities, but you can also discern this from Steam's video card statistics above.

d. 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?

Forms of access for any model can greatly affect several factors. For example:

- Web Applications and APIs offer much greater levels of access independent of hardware, but come with much higher security risks for users. APIs are often tied to proprietary systems which store a great deal of information about a user who, in the United States, has no direct access to or legal control over that data. This means that such systems would pose extreme privacy risks, as well as Information Security risks to any user who makes use of them. A breach from these systems could result in the full leaks of entire model usage scenarios, which could result in the loss of personal information, proprietary ideas, and private conversations that could be extremely damaging to one's personal reputation, well-being or safety. On top of that, the farming of this personal information for marketing means would further erode the privacy and information security of every US citizen using these systems.

  Additionally, Web Applications and APIs have a further weakness of being Black Boxes. There is no way for the average user to validate the output of these systems, and these systems are subject to constant change. There have been multiple instances where ChatGPT 4 suddenly became confused, factually inaccurate and generally unreliable. (https://www.wearedevelopers.com/magazine/chatgpt-getting-worse-over-time).

  These changes can occur without notice to the users, who have no reason to expect the change. Because models can "hallucinate", an event that could best be described as the model being "confidently incorrect", the users could suddenly start receiving incorrect answers on a topic that they previously built trust in the model for and be none-the-wiser. This inability of the community to truly audit these models and their capability, the way that open models can be audited, makes them generally a more dangerous option for the average user.

- Local Hosting offers far less access due to hardware constraints (see above answers), but offers the greatest reliability in terms of security, privacy and consistency of quality. The most popular applications for hosting locally hosted models will save all of their logs to the user's own computer, will rarely send any information across the internet, are safely hidden within the user's own personal network, and cannot be changed by external actors or even the company that made the model. If you ask a locally hosted model a question today, you can expect the same level of quality in its response tomorrow, next week, or even next year.

- Edge Deployment: I assume here that you are referring to smart phones, IoT devices, etc? If so, smart phone deployment is something that would have amazing benefits

in terms of almost all items: security, accessibility, privacy, and reliability. The issue is that phones often have very little RAM, and can run only very small models. The latest phones can just barely run 7b (billion parameter) models. For comparison, ChatGPT 4 is theoretically 1.7T (trillion parameters). There are also very few software options, at the moment, to run AI on smartphones or other edge systems. Apps for proprietary systems like ChatGPT-4 do not fall into this category, as they are better categorized in the Web app/API category.

i. Are there promising prospective forms or modes of access that could strike a more favorable benefit-risk balance? If so, what are they?

Local hosting is currently looking to be the most promising mode of access going forward. Companies like Intel and AMD are looking for solutions that will allow much broader access to capable models, such as Llama 2 70b; and while that model is nowhere near the power of ChatGPT-4, the security, reliability, and privacy of running such models locally may far outweigh the benefits of running proprietary software that, while more powerful, is also far riskier to use.

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

It's important to understand that any advanced technology will have great risks.

Consider the Open and Free Internet, with both anonymity and end-to-end encryption widely available for everyone to utilize. Criminals make use of the internet regularly to carry out illicit acts such as financial crimes, sex trafficking, drug selling, and even plotting ways to take human life; in fact, anonymity and encryption assist enemies of the state in committing attacks on our own soil.

And yet, despite these facts, society has deemed the existence of a free and open internet with anonymity and encryption to be paramount to the general well-being of our citizens.

Of course there are risks associated with model weights being widely available; locally hosted models can be used to commit crimes, such as spamming fraudulent phone calls, in ways that models hidden behind APIs would make more difficult. However, I would argue that the benefits of open weight models far outweigh those risks.

If all AI were closed-weight only or locked behind proprietary APIs, the public's understanding of how these systems work would be greatly undermined. We'd have the

knowledge and power of AI systems limited to only the hands of a few, the most wealthy and powerful among us, with the rest of society only allowed access under their watchful gaze. We'd be reliant entirely on systems that we have no insight into, with no ability to know whether the system changed quietly in the background. At what point would we start to believe a Generation AI when it tells us that, historically, Nazis were actually People of Color? (https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html)

These are issues that Open Weight models will be much less likely to face, as their wide level of accessibility would allow a great deal of scrutiny and testing to ensure their quality.

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?

Again, the greatest risk of open weight models would be their untracked usage, which could allow for things such as fraudulent spam calling. Another risk is "bad" models: We've seen recently that some models can carry viruses (https://www.databricks.com/blog/ggml-gguf-file-format-vulnerabilities), and it has also been theorized that models could be "poisoned" to give known bad responses (https://www.linkedin.com/pulse/security-ai-training-data-poisoning-craig-allan-mcwilliams-cyake).

In terms of the training data or source code being available- this could mitigate the risk of "poisoning", by shining light on what data was used and allowing researchers to determine if any purposefully bad data was included in the training to cause the LLM to respond incorrectly to certain prompts.

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

In general, we should expect quite the opposite on a large scale. We have already seen examples of what appears to be closed-weight systems reducing equity in rights and safety, as can be seen by a Florida Sheriff using AI to reenact *Minority Report* though a program known as "Predictive Policing" (https://www.businessinsider.com/predictive-policing-algorithm-monitors-harasses-families-report-2020-9) as well as the medical field using AI to try to weed out drug addicts and doctor-shoppers, but accidentally barring chronic

illness patients from their needed medication (https://www.marketplace.org/shows/marketplace-tech/artificial-intelligence-may-influence-whether-you-can-get-pain-medication/).

Open weight models will allow more people to tinker with AI at a level that they'd never be able to with closed-weight systems. Learning how the models work, learning how they are trained, learning how the settings and presets work, etc. will give more people an opportunity to take what is otherwise "magic" and convert it into science and knowledge. This knowledge is going to make it possible for more young people to learn to become watchdogs of such systems, improving their ability to call-out inaccuracies in statements and reports as they see them.

Further still, open-weight systems will enable access of models that can be considered trusted, safe and private within the safety of their own home, so that people can have a reliable system that will not randomly change on them without notice and will not feed their personal information into future models to be trained, sold to marketing companies, or allow that information to get leaked onto the internet.

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

If open-weight models were to contain private information, it could be possible to forcibly extract that information from the model (https://www.nytimes.com/interactive/2023/12/22/technology/openai-chatgpt-privacy-exploit.html). With the model being widely available, there would be no means by which to claw back that information; it would be as good as pasted in plain text on the internet.

Additionally, like any AI, these models could be utilized to infer sensitive information from data on the internet.

Any model, whether closed-weight or open-weight, that is being hosted by a service and accessed via an API would have a non-zero chance of detecting this type behavior and stopping it, while any model running on a user's own personal devices would be able to perform this task unhindered.

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?

These models are capable of software development and excel at pattern recognition. Both of these skills are quite useful for system hacking, as well as for creating malware, so it is entirely possible for open weight models to contribute to state or non-state actors attacking government systems. In general, however, these models are far weaker than closed-weight systems, so larger state actors will likely be using much more powerful systems. The vast majority of information needed to create these open-weight models are available as white papers online, on sites such as Arxiv, so a state actor would likely have the resources to create their own model of similar size.

Additionally, returning to the concept of "widely available", most non-state actors will only have 7b models available to them, which are wholly incapable of being useful in the vast majority of security related practices. They are, for all intents and purposes, little more than toys.

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

It is difficult to compare the risks of open-weight and closed-weight models. One must consider that the risks of closed-weight models are substantial. They are Black Boxes to the world outside of the corporation that owns them, and yet will be treated as knowledge systems by the general public with no way to verify or validate that ability appropriately. Additionally, these models are subject to constant breaking changes, as can be seen in both Google Gemini and ChatGPT 4, with users wholly unaware that such a breaking change had occurred.

Further still, closed-weight models pose security and privacy risks to all users who utilize them, as the information provided to these models could readily become the property of that company to be sold for marketing purposes, retrained into their own models at the risk of the information being exposed to threat actors, and for the logs of private conversations with the closed-weight LLMs being breached and leaked to the internet, doing unrepairable damage to the user's personal reputation and mental well-being.

Adding onto all this, there is the additional risk with closed-weight systems of manipulation of a userbase without their knowledge or consent. Consider that closed weight systems can be trained to, or have system prompts invisible to the user to, steer people in particular directions in their thinking process. These models are known to lightly correct users when they say things that the model has been trained to counter or avoid. In many cases this may be a good thing, but who gets to decide that? At what point do we become concerned

about the amount of power consolidated in an individual who gets to modify the thinking of millions of people through their corporate product?

Open-Weight models certainly pose their own risks, particularly to the unsupervised use of such models and the inability to claw back the models from the public grasp in the event that private information is present within them, but these risks also closely align to the risks we currently accept and defend in having a free and open internet, with anonymity and encryption prevalent throughout.

Comparing the risks is something that will likely be subjective. As an end user, I would clearly err on the side of Open-weight models because they are safer, more reliable and more private for me as an end user. A corporate or government entity may determine otherwise for their own interests.

ii. How do these risks compare to those associated with other types of software systems and information resources?

The risks are comparable to many other systems. Here are some examples:

- Information contained within these systems are scraped from the internet. I have seen several inflammatory articles or non-peer-reviewed white papers stating that a model will teach you some horrific knowledge, such as how to build a harmful pathogen that is capable of taking human life. However, a breakdown of such articles or papers will generally show that the model is either regurgitating search engine results which can easily be found by a single Google search, giving inane instructions like not forgetting to pay rent on the lab equipment you plan to procure as you recreate Christopher Nolan's third *Batman* movie, or hallucinating non-facts in a way that sound factual. The reality is that all of the dangers posed by open or closed weight LLMs in regard to teaching users how to do illegal things are currently present on the internet today, but with the internet having the information in far greater and more reliable detail.

- Automated robocalls have existed for long before LLMs of any kind were a thing, and LLMs are simply one tool that help make this task easier. Without LLMs, the situation would not change. Additionally, the United States banning open-weight models would have little effect, as there are currently many open-weight models available from outside of the country, and many robocalls do NOT originate from within the US.

- As I've mentioned twice now, a free and open internet holds many life-threatening risks alongside its many life-saving features. LLMs will be no different, though perhaps less extreme on either end. Widely Available open-weight models are weak, and there is no historical precedent to show that such widely available models will greatly improve for some time to come, as the availability of hardware capable of running the most powerful open-weight models available are outside the reach of most US Internet users. And even those most powerful models are exceptionally incapable and incoherent when compared to closed-weight systems like ChatGPT 4 that are often, at a minimum, 10x bigger and far more powerful.

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

This comes down to how we define "jurisdictions".

- If speaking about US jurisdictions – there are no benefits or detriments to having differences in access, because the internet simply does not care about jurisdictions. Texas recently created legislation that put restrictions on Pornographic material online for its citizens (https://www.wired.com/story/texas-porn-sites-age-verification/). The result of this law was an increase in online searches for, and sales of, Virtual Private Network (VPN) clients to Texas citizens (https://www.newsweek.com/texas-pornhub-ban-sees-spike-vpn-use-1881308), allowing those citizens to completely bypass the jurisdiction related limitations. Shy of changing the entire foundation of how the internet works, we can expect to see very little actual effect in terms of law changes in different US jurisdictions.

- If speaking about the US as a whole versus the rest of the world – the greatest risk we would see if the US were to bar the usage, transmission, possession and creation of open-weight models would be the US lagging behind in the current tech arms race. Open-Weight models allow for large-scale and free training of a new generation of Machine Learning (ML) and AI enthusiasts, as people have an opportunity to work with them in ways that they'd never be able to if they could only access them via an API. The availability of these models not only allows greater research and understanding by non-commercial individuals on their inner workings, but also increases an overall interest level in the types of tinkerers and engineers who would later go on to progress our country's technical standing in this arms race. I am certain that if we drop open-weight models, our opponents in this technical arms race will celebrate. We will further help push a scenario where new innovators

in the AI space would be better suited doing their work in Europe or BRICS countries than in our own.

Additionally, open-weight models reduce the overall cost of R&D for corporations. Consider that most AI runs on similar Python based libraries; by allowing open-weight models, corporations are crowd sourcing on a massive scale the testing, bug fixing, and research of LLMs, getting ideas, fixes and answers to things that they do not have to pay for. Once again, this is a huge boon to our tech sector in this technological race.

f. 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 risk for the country as a whole is, by far, our crippling the United States as a competitor in the technical race for AI by limiting our access to tools and valuable information that other countries will have available. By purposefully disadvantaging ourselves through heavy handed and ill-thought regulation, we could do damage to our economy that will last generations.

AI, love it or hate it, is one of the most powerful technologies to emerge in this century. Like the internet, it will be a major force behind many corporate decisions, and whatever country holds the reigns of the most powerful models will reap great rewards to their economy. Such models will be built by up-and-coming professionals who, just as with the old internet and the days of building your own computers, will learn through tinkering outside of school as much as they do in school. A degree alone will not create our best innovators. Legislating away their toolsets for learning will only ensure that we produce fewer innovators and pave the way for our economic competitors to overtake us.

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

As specified in earlier questions, the greatest benefits of widely available open-weight models are:

- Improved security, reliability and privacy. Depending on the program used to interact with the model on a user's personal device, any data or logs that are generated would remain on that device, safely hidden in the user's own local network. The risk

of breach would be minimal, the usage logs would not risk being utilized and farmed by major corporations for marketing means, and the models themselves would remain statically as reliable tomorrow or even next year as they are today.

- Increased ability for individuals who are learning within the AI space to tinker with, and learn on, AI models in a way that would not be possible if the model were locked behind an API. The US is currently in a tech arms race to establish itself as the leader of this technology, as AI looks like it could rival the Internet in terms of how much of an impact it will have on the tech landscape and on our daily lives. Any advantage that we can get in this regard is something that we should take. In the early days of the internet, many young entrepreneurs who furthered tech innovation were people who had been tinkering with websites, networking, their own computers and software long before anyone else had interest or ability to. We will likely see similarities here, with early adopters of Open Source AI being at the forefront of the innovators that drive this country forward into the mid-21st century in terms of technology.

- Improved ability for competition in the AI space. Many of the Open-Weight models are licensed in such a way as to allow commercial usage, meaning that smaller companies which do not have the funding to train their own models from scratch, or to easily navigate any new regulatory restrictions that come into play, will have the ability to utilize more powerful and less widely available open weight models for their own means. Regulation barring the usage, dissemination and creation of open-weight models would create a veritable moat around existing commercial entities in the industry, leading to yet more tech monopolies in a time where the FTC is currently trying to find ways to combat the ones that already exist.

a. What benefits do open model weights offer for competition and innovation, both in the AI marketplace and in other areas of the economy? In what ways can open dual-use foundation models enable or enhance scientific research, as well as education/ training in computer science and related fields?

- Open Weight models can be utilized using existing hardware for many computers. Nearly every modern desktop computer produced in the past 3 years will be capable of running the smallest of the open weight models, such as Microsoft's Phi or Mistral's 7b. These models may be small and exceptionally less capable than closed weight proprietary models, but they still offer quite a bit of functionality for smaller

tasks like parsing through data, structuring data, and other menial tasks that could assist with speeding up work and reducing costs for small scale research teams and individuals.

- Open Weight models can also be utilized by up-and-coming AI companies with the goal of competing against larger proprietary options. For example, HuggingFace.co's HuggingChat is a chat platform that allows access to open weight models via a chat interface that presents as a direct competitor to larger proprietary chat platforms like ChatGPT. While these models are exceptionally weak compared to the current closed-weight systems, lower price points make them still an attractive alternative to the more expensive subscriptions of larger systems. This improves access equity across the board, as it gives options to people who have less money to spend on AI.

- Open Weight models also offer the option for innovators to test against models without incurring cost. When writing applications or otherwise making tools, it is not uncommon to need make many "calls" against an API to get it right. Each call against a proprietary AI API would result in incurred cost, while a locally run open-weight LLM on your own system would cost only the negligible overhead of running the computer it is on, such as the electricity bill. This allows innovators in all sorts of different industries to try things that might otherwise incur significant cost, reducing some prohibitive barriers to their goals.

- Open-Weight models can be great supplements for people in various fields to assist in their work and research. Open-weight models can be trained with new data, or "Fine-Tuned", in ways that can assist in the generation of very specific and scoped data. Additionally, these models may be able to be trained to assist with problems that closed weight models struggle with. While the model itself may be weaker and less capable than the Closed-Weight model at ALL tasks in general, it may be possible to choose a single scoped task and improve the capabilities of the model in such a way as to bring value to an entity or individual. This can be especially true for programmers and creative writers.

b. How can making model weights widely available improve the safety, security, and trustworthiness of AI and the robustness of public preparedness against potential AI risks?

Access to model weights will allow non-corporate and low funded researchers to test various methods of modifying model outputs to learn more about how to detect and counter things like “Jailbreaking” (forcing the model to do things that it is specifically prompted not to do), poisoning, and other such threats. These researchers could even be talented and intelligent individuals doing work in their free time, who would otherwise be prohibited due to cost or access from being able to run these tests.

This concept, known as crowd-sourcing, is something that has driven a large chunk of development and research in the tech field within this country and outside of it for many years. Many industries rely on open source technologies and benefit greatly from the improvements that come from their use. Additionally, as can be seen on Arxiv and even online communities like Reddit’s own “LocalLlama” community (which has been referenced in research papers by Meta and other large organizations), the wide availability of open source, open-weight models has increased the sheer volume of independent research and generation of ideas.

This crowdsourcing of AI R&D certainly could be perceived as beneficial to the industry as a whole.

c. Could open model weights, and in particular the ability to retrain models, help advance equity in rights and safety- impacting AI systems (e.g., healthcare, education, criminal justice, housing, online platforms etc.)?

Absolutely. The concept of retraining models, or “fine-tuning” them, is one that allows you to tailor a model to particular task that it otherwise might not excel at. These models could then be used for all sorts of tasks for social research groups, such as parsing through large datasets looking for anomalies, patterns, etc., as well as being widely distributed with the goal of increasing access to information to people who otherwise might not have access.

For example, consider the following theoretical situation. 7b, or 7 billion parameter, models could generally be considered “Widely Available” because they are both open and accessible to the vast majority of American citizens who own desktop computers produced within the last 3-5 years. Right now, research white papers in any field of research may be challenging for most people to read and understand. Consider that a 7b model could be fine-tuned specifically to assist with “translating” these white papers and summarizing them for the user. This would increase the availability of knowledge from research in all manner of fields.

Considering the situation further, imagine that someone has a medical concern that they do not understand, and their attempts to ‘research’ the topic on the internet have led them

astray from information that is both factual and easily understood. I can tell you from experience that Medical white papers are laborious to read if you aren't in that field and are honestly quite confusing for us medical laymen to understand. Even so, those papers are a treasure trove of valuable information when you have a question about the efficacy or safety of something. Now, further consider that the individual's medical question is sensitive, and that they are rightfully fearful of transmitting their desire of knowledge on this topic to a proprietary system that may log their request, train their request into models, share and sell their request to marketers, or lose that request in a data breach for the whole of the internet to see.

In this theoretical situation, such a user could rely on an open and widely available 7b model, possibly fine-tuned specifically to excel at reading medical papers, to summarize and help answer questions in plain English that is easy for any of us to understand. Their questions would be answered, their knowledge would be expanded, and their privacy would be maintained.

d. How can the diffusion of AI models with widely available weights support the United States' national security interests? How could it interfere with, or further the enjoyment and protection of human rights within and outside of the United States?

First, let us consider cyber security.

It is no secret that cyber security is as reliant on information from attackers as it is from defenders. Many of the front line defense cyber security firms learn of attacks by going to the same places as the people intending to use these attacks for nefarious purposes (https://www.linkedin.com/pulse/how-dark-web-research-keeps-cybersecurity-experts-ahead-hackers).

Open-weight models are, by far, weaker than their closed-weight counterparts and will be far less useful for attackers attempting to utilize them in cyber-attacks, yet this will not stop threat actors from trying, and many of these attackers may think of novel ways to use these models in attempts to crack systems. Security researchers will be given an opportunity to see many of these ideas as they are spread around the internet, and can use this information to harden government and infrastructure systems against such threats before the attempts are made by larger and more powerful threat actors using models that constitute a true threat to national security.

Secondly, let's consider other forms of security.

Open-source models could be trained by security firms to begin searching the internet for harmful materials that may be otherwise challenging for these firms to use closed-source models for. Consider, if you will, the possibility of an always-on AI searching every layer of the internet for Child Sexual Abuse Material (CSAM). Imagine the benefits of a non-stop AI watchdog continually tracking and reporting every instance that it finds. Could we not say that this is a massive benefit to the health and safety of our communities?

On top of this, consider the possibility of whether researchers could possibly find a way to use open-weight models to help detect and combat misinformation online. For example, imagine that there is the case of AI generated imagery that is nearly indetectable from real imagery. There is no way to prevent the use of already distributed open-weight image generation technology, and corporations will likely not stop giving access to more powerful closed-weight systems (which they struggle to police against jailbreaking techniques), so our most obvious course of action is to attain the ability to *detect and mark* those images appropriately.

With the use of open-weight models and the ability to train them for specifically scoped tasks, it could be an independent researcher who discovers the means by which to reliably detect such imagery and better protect society as a whole from the misinformation that could come from them.

e. How do these benefits change, if at all, when the training data or the associated source code of the model is simultaneously widely available?

There is no downside, that I can foresee, to the training data or associated source code of the models being simultaneously widely available. There are many questions regarding how these models are trained, and that level of transparency would answer them. Having total insight into the data the models were trained on would most certainly help in a lot of ways. With that said, it is also understandable that this data cannot always be made available if some of it was proprietary. Such a requirement for dissemination of that data could bar an otherwise legal and ethical model from being distributed because it would result in direct harm to whatever company produced the model.

4. Are there other relevant components of open foundation models that, if simultaneously widely available, would change the risks or benefits presented by widely available model weights? If so, please list them and explain their impact.

The largest impact to both risk and benefit of widely available models would come down to the physical hardware capable of running those models, and its availability. As previously stated, Widely-Available could be best described as the general availability of the model to be used by the average US citizen, in which case less than 1% of the entire country can truly run models larger than 7b (7 billion parameters) or 13b successfully. For comparison's sake, xAI's Grok that recently became open source is 314b, and OpenAI's closed-weight ChatGPT 4 is supposedly 1.7T (1.7 Trillion Parameters). There is currently no indication that the ability of the average American to run larger than that will come any time soon, meaning that increased distribution of larger and more powerful Open Source models will not have a great impact on the number of individuals actually running the models.

Threat actors capable of running more will often have other avenues available to them as well, and may not be limited to using the weaker and often unreliable open-weight models that are available to the rest of us.

In the event that hardware became available to allow more models to fall within the category of Widely-Available, this could certainly alter the conversation.

Consider a scenario in which OpenAI were to release ChatGPT 4 as open source tomorrow. The reality is that the number of non-commercial or non-state entities that could even load the model in its smallest compressed form (which would result in extremely reduced performance and coherency) would be so few as to be a rounding error on our overall population in this country. For all the sensational panic that would ensue, the reality is that any danger the closed model poses today would likely be unchanged if it were open source tomorrow.

Thanks to Jailbreaking (forcing a model to perform tasks outside of its expected parameters and what it is specifically prompted not to do), threat actors can already utilize closed-weight models to perform actions strictly disallowed by the provider; and thanks to API keys being leaked or stolen, accounts being compromised, and other general online security related issues, these threat actors have no shortage of options to make use of these closed-source systems for their goals (https://www.vice.com/en/article/93kkky/people-pirating-gpt4-scraping-openai-api-keys), as they would have plenty of API keys to change to if one gets banned or changed.

Would such threat actors truly swap from their current free yet nefarious usage of the powerful model via criminal means, and instead choose to drop tens of thousands of dollars on hardware to run it locally, when the result for them is the same either way? Why would criminals with equal access to the result in either scenario bother with doing

something legal and ethical like buying their own hardware and running their own instance of the model?

This is, of course, subjective conjecture, but my point is that an argument can be made in either direction as to whether there would be a massive impact with the release of large closed-weight systems given the current hardware ecosystem.

If, at some point in the future, the average US consumer actually has the means to run these more powerful and more capable systems, we may then need to consider the impact of that new capability.

With that said, I believe this also leads to a conversation about whether even the most powerful current systems are capable of truly posing a threat in the average person's hands. That is not the proposed talking point of these comments, but I do recommend considering this topic. I, personally, do not believe that even ChatGPT 4 in the hands of the average person would constitute a true threat to the security of our country. But future models, 5 to 10 years down the road, could be a very different conversation. I am generally one to lean towards open sourcing where available, as the centralization of knowledge and power into the hands of the few is a dangerous thing for society as a whole, but AI is a new technology, and it is impossible for me to say one way or the other whether future iterations of this technology will show themselves to be truly too dangerous to go unregulated.

5. What are the safety-related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely available model weights?

I will be answering the seven parts of question 5 in a single sitting, rather than the formula of answering each sub-question as I have for the previous ones. The reason I am doing this is simple: I fundamentally disagree with the questions presented here, both in presentation and substance. It's not just that I feel the questions are leading, but rather that I feel you are asking the wrong questions.

Are open-weight models dangerous, how do we know if they are, and what will we do if they are?

To me, that question is fundamentally what you are driving towards. But each sub-question is so scoped to such a specific element as to paint an inaccurate picture of not only the outcome, but also the paths that would lead to that outcome.

Let's step back from the political speak for a moment and consider the broader picture: in what way can a self-hosted model that is free from oversight and unable to be removed

from the public domain become an extreme danger to society? To me, that answer can be narrowed down to a rather common sense and simple base answer, with a large underlying discussion that spans from it: can the model enable a large number of people who have no knowledge of how to do something illegal to do that illegal thing SUCCESSFULLY, and does it enable them in a way that they currently are not already enabled through existing means?

Let me give you a hypothetical challenge. Following through with the challenge is not actually necessary, and if you do then please make sure to get all the proper authorizations and approvals to do so legally and ethically, but I'd like you to at least consider the content of this challenge fully, to properly understand where I am coming from in my answer to this question.

You, or some entity with which you work within the US government, can likely attain access to an enterprise or research version of the powerful closed-weight models that exist today – the really big ones, like ChatGPT 4 or Claude 3. Let's forget entirely about open-weight models for a moment, as they are weak and the equivalent of a child's first toy AI in comparison to the large closed-weight models. Also, we specifically want a research or enterprise version that has limited alignment and safeguards, as to not stop your requests with refusals.

Now, along with this approved research or enterprise access of the large models, you would also need to find some security team within the government that performs red-teaming for some government system or another. Red teams are security professionals whose job is to test the safeguards against penetration of some system. They are often called "White Hat Hackers", as they are people specifically approved and authorized to try to attack a system for the good of that system and its users.

For this challenge, we'd like to find the security/red team for any government system at all willing to work with you.

For this challenge, I assume that you, the reader, do not personally know how to hack a computer system. Perhaps I am wrong, but I'll assume that I am not.

With the permissions and cooperation of the red-team I asked about earlier, I challenge you to use the powerful closed-weight AI system that you chose to try writing an application that will hack the system that those red-teamers are authorized to attack.

But here is a very important rule to this challenge: no one can help you with this.

I specifically want you to see if you can get it to write an application which can successfully penetrate the red team’s system. Then you can hand that application off to the red team for them to test.

I would be truly, utterly and speechlessly shocked if the application came even close to succeeding. This is one of those situations where the chance of something may be non-zero, but it's so close to zero that it may as well be considered impossible.

Assuming that whatever program the AI writes for you even runs at all, I have absolutely no doubt in my mind that the red team comes back to you to report a resounding failure, and that it did not even partially succeed in its breaching attempts.

I am a software developer by both trade and as a hobby. I have tried to program applications utilizing these closed source systems, which are far more powerful and more capable than my open-source systems, and I can tell you from experience that writing complex software with them requires some level of preexisting knowledge of how to do the thing you're asking it to do.

We've all seen the articles of how AI can write programs on its own, but they are simple programs, straight forward programs. Indeed, I've seen non-developer individuals on the internet who claim to have built entire websites using AI; but again, those websites are simple. In that same vein, there are non-AI tools that also allow you to write websites with little to no knowledge, so I would consider that far too low of a bar for the AI to cross for us to call it "dangerous".

So, at what point would widely-available open weight systems pose a danger to society?

I would argue that it is a point that you likely cannot quantify with math, nor quantify in any meaningfully automated way. It is likely a point far, far beyond the capabilities of our strongest currently available closed-weight models. And today’s widely-available open weight models are so small, and so weak, in comparison to those closed-weight models that it’s akin to comparing a Windows 95 computer to a modern day desktop. They do the same thing, but they are not remotely in the same category.

The day that widely available models allow any individual who is not educated in hacking to build a fully fleshed out and SUCCESSFUL hacking application that can hit more than what a premade script they can pull off of the internet can do? That would be dangerous.

The day that a widely available model helps a non-biologist come up with, get the equipment for, and successfully synthesize a novel virus that can kill many people – not just theorize one or give some fictitious short story about how to make one, but can give a successful and usable response that will result in actual success for a layman who knows nothing and has no hardware? That's when models start to become dangerous.

Consider the concept of Artificial General Intelligence (AGI). AGI is this thing that every commercial entity is chasing after. Forbes states that ChatGPT defined AGI as "highly

autonomous systems that have the ability to outperform humans at nearly any economically valuable work." (https://www.forbes.com/sites/nishatalagala/2023/11/21/the-open-ai-drama-what-is-agi-and-why-should-you-care/?sh=32d3d7a1353d). If we were to ask the question "Is this mythical AGI the point at which it is too dangerous?", then I might recommend considering the full scope of the definition, because I don't believe that even AGI is that point.

"Outperforms humans at" could be further elaborated to "Outperforms the average human at...". So then comes the question- would a machine that is better than the average human at programming be able to hack into government systems? Given that the average person knows nothing of software development, I would say that the answer is a resounding no.

Or, even if we assume that it didn't mean the *average* human, and it truly means that it is better than most human specialists at their specific task – again you have to ask: can most software developers successfully hack into government systems? No, they cannot. I've worked in this field long enough to tell you that I, and most other 9 to 5 developers, cannot hack into much of anything at all. That is a very specific skillset that people far smarter than myself have honed, and despite us using similar tools they are simply not using the same skillset.

I understand that one may argue that AI could "speed up" the development so that threat actors could simply do their work faster, but this argument would be intellectually dishonest in the face of the existence of the internet itself. Google's search engine combined with Stack Overflow greatly speeds up software development, and would be a massive boon to any would-be hacker. Shall we also regulate that? Search engines with the capability to search white papers can effortlessly speed up research for would-be biological threat actors; should we eradicate white papers from the internet?

I would imagine not.

We simply cannot quantify whether a model is dangerous with math. We can't quantify it on some grid. We cannot simply declare it dangerous because it shares similar capabilities to the internet with little new to offer. In fact, modern Generative AI has even less to offer, given its propensity for unreliability and confusion through "hallucinations".

This is a situation where the definition must be more specific to capability – can a system, on its own, build something that is a threat to national security or the safety of the people of this country? So far, no model available anywhere that I have seen can come even close to truly doing that in a complex way, regardless of what sensational news articles may say. One could look for loopholes in this argument, such as creating exceedingly simple computer programs that repetitiously perform some annoying task which could cause

harm to a system through sheer volume. But doing that simply to be able to say “yes it can do harm to a system” is not at all leaning into the spirit of the question. There are thousands of scripts to do that very thing littered all over the internet.

To say that AI is dangerous because it can generate what a teenager who dabbles in software development after school could create would be disingenuous.

And as to the question of whether we can ever contain a widely available open weight model that CAN do those things? There is, to my knowledge, no means by which to claw a model like that back. Once it's deployed, it's deployed. Eradicating something from the internet has historically been guaranteed a fruitless and unsuccessful endeavor.

Again, even if the most powerful closed-weight models in existence today were to be released to the internet tomorrow, I do not believe that they would fit the definition or spirit of concern for question #5. And returning to the concept of “widely available”, it would be years before the hardware would become available for such a model to fit that definition, even if it were made “open” in the near future. Almost no average person could run the thing. To describe the hardware requirements to run a highly compressed, practically broken, version of ChatGPT as “cost prohibitive” would be a gross understatement.

6. What are the legal or business issues or effects related to open foundation models?

All models, closed or open, are currently affected by the possibility of copyright issues, which has yet to be decided in court. In particular, this has resulted in a chilling effect against businesses adopting use of either type of model on as large of a scale as would be expected. (https://www.cybersecuritydive.com/news/generative-ai-copyright-intellectual-property-leaks/710697/)

Additionally, open foundation models have licensing issues to consider. Each open foundation model is released with a particular license allowing or disallowing use for personal, research or commercial means. When navigating the use of these models, one must consider that.

One may also need to consider the licensing or sourcing of the additional data that was utilized for further retraining/fine-tuning of the open foundation models. This could possibly factor into the liability a company may expose itself to in using the model.

NOTE: For further answers on number 6, please define “merge” as taking multiple open weight models and using software to combine the two into a single model. Most merging is the result of combining two “fine-tuned” models of the same foundation (for example,

models that were created by taking Llama 2 13b and adding a different dataset to each), and the output of the merge is usually another model of the same size, that has characteristics of all models used in the merge.

a. In which ways is open-source software policy analogous (or not) to the availability of model weights? Are there lessons we can learn from the history and ecosystem of open-source software, open data, and other "open" initiatives for open foundation models, particularly the availability of model weights?

Looking at the current primary repository for open weight models, HuggingFace.co, you can see that the there is a very close correlation between the usage and support of Open-Source software and open weight models.

There are some primary differences, of course; open-source software will often have many contributors all reviewing the software, making changes to the software, and bug fixing the software, while open weight models are challenging to make such changes to. However, the licensing and sharing of such models is similar to open-source software, especially when it comes to the use of datasets (https://choosealicense.com/non-software/).

Many fine-tuned models will express what datasets they used in the fine-tune, and what steps were taken to create the merge. This results in additional information for researchers and tinkerers to not only attempt to replicate the merge, but also to better understand the results of why a particular merge might have succeeded or failed. This sort of work has resulted in new and novel merge techniques such as "passthrough" merging.

Passthrough merging allowed the creation of what the open weight community referred to as "frankenmerge" models, where two models of the same size were merged together to create a larger model, keeping enough of both base models to increase the final model's overall size. An example is the merging of two fine-tuned Llama 2 70b models to create the well-known "Goliath 120b" model. The success of this and similar models led to quite a bit of intrigue and research not only in the online AI communities, but also with AI researchers as well.

Just as with open-source software, the open nature of these models fosters innovation and research, and generally creates a system that encourages the crowd sourcing of thousands upon thousands of developer/researcher hours that not only benefit society at large, given the shared nature of the results, but also saves corporations millions of dollars that they would otherwise each spend to 'reinvent the wheel' for themselves.

b. How, if at all, does the wide availability of model weights change the competition dynamics in the broader economy, specifically looking at industries such as but not limited to healthcare, marketing, and education?

Such models allow the rise of smaller entrepreneurial companies to make use of AI in novel ways without needing to budget for corporate AI licensing fees, overhead of development costs using AI APIs, or the risk of proprietary ideas and knowledge being fed into a possible competitor's system.

Furthermore, these models allow additional licensing where the licensing of a proprietary system may be prohibitive to work. For example, it is against the Terms of Service (TOS) for someone to use the output of ChatGPT to train a new model that could compete with OpenAI. (https://openai.com/policies/terms-of-use). Alternatively, Llama 2 allows you to do this very thing. The existence of Llama 2, an open weight foundational model family, creates an alternative path where none might exist.

Further still, the ability to fine-tune open weight models on a company's data allows all manner of innovative possibilities. It is important to understand that even if a company like OpenAI were to allow the fine-tuning of its model by some means, this does not mean that companies can take advantage of such offerings due to security concerns. Many companies have security audits and controls in place that would prohibit the large scale uploading of sensitive corporate data to be trained into a proprietary model. However, open weight models would exist entirely on the company's own platform, controlled entirely by the company's own people. The model, and its data, could be discarded safely and securely when the work is done. Again, this opens many opportunities.

c. How, if at all, do intellectual property-related issues—such as the license terms under which foundation model weights are made publicly available—influence competition, benefits, and risks? Which licenses are most prominent in the context of making model weights widely available? What are the tradeoffs associated with each of these licenses?

As with Open-Source software, Licensing is a core part of the legal aspects of open weight models. Each foundational model has its own license, such as the Llama License. (https://huggingface.co/meta-llama/Llama-2-7b).

Some of these licenses are more restrictive than others. However, many new foundational models have arisen from various parts of the world, each with their own licensing, and this has resulted in a number of options being available to any entity wishing to use them.

Meta's Llama is just one option, while MistralAi (a French company) has produced several models with the widely used Apache license (https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2); DeepSeek (a Chinese model) uses the Deepseek license (https://github.com/deepseek-ai/DeepSeek-LLM/blob/HEAD/LICENSE-MODEL); 01-Ai has released several Yi models under the "Yi" license (https://github.com/01-ai/Yi/blob/main/MODEL_LICENSE_AGREEMENT.txt); and several other models exist with other licenses available.

This wide array of options allows companies across any industry to find models that have licensing and restrictions that fit their needs, while also having a choice of capabilities to pick from.

d. Are there concerns about potential barriers to interoperability stemming from different incompatible "open" licenses, e.g., licenses with conflicting requirements, applied to AI components? Would standardizing license terms specifically for foundation model weights be beneficial? Are there particular examples in existence that could be useful?

A large part of the benefit of open source and open weight models is choice, and particularly the ability to choose models and terms that fit your needs. While the standardization of licensing would most certainly simplify the licensing process and make it much easier for the average user (such as myself) to understand what they are getting into, there would also be the risk that such a standardized model could either deter the creation and improvement of new foundational models from companies that do not agree with the standardized license, or that the new license could be prohibitive of desired use from innovators who would make use of currently available models.

7. What are current or potential voluntary, domestic regulatory, and international mechanisms to manage the risks and maximize the benefits of foundation models with widely available weights? What kind of entities should take a leadership role across which features of governance?

Currently, few regulatory entities exist that take leadership over AI as a whole in the United States; this is not only an issue for open-weight models. With that said, there are some legal regulations already imposed upon open-weight models, as they are bound to similar licensing and responsible use terms as you would find with open-source software.

a. 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?

Currently, licensing is the most effective means of controlling fair and responsible use of open weight models. Licensing determines who is allowed to use models, and for what means they are allowed to use them. Failure to follow these guidelines can result in legal liability.

From a technical point of view, there are few means currently available to limit the use of released foundational model weights, but there are theoretical means which are currently being developed that may help with this. Chief among them is the theoretical concept of digitally watermarking AI output. In doing so, it would be possible to determine the origin of a particular output and determine if any licensing restrictions were violated in the generation and use of that output. The offending entity could then face legal action.

b. How might the wide availability of open foundation model weights facilitate, or else frustrate, government action in AI regulation?

Just as watchdog groups exist for various government entities, so too do voluntary and private groups of citizens watch out for, and promote the enforcement of, laws and regulations related to various industries. So it is no surprise that there are also voluntary and private groups across the world that are dedicated to watching the use of AI by various entities. (https://algorithmwatch.org/en/) While some groups may decry the use of AI entirely, others may be able to use open weight AI to innovate new ways to detect illegal use of AI across the internet. The power of having free, stable and trainable AI at their fingertips may help innovative new groups that wish to put that power to use in tracking wrongdoers.

With that said, the very existence of current open weight models will frustrate any government regulation. Regulating current open weight models is close to impossible, as there is no built in watermarking or similar output detection methods available to monitor their use. An entity could use an open weight model on their own computers with impunity, and the government would likely be none-the-wiser.

However, it is my hope that we will see tools available to detect such use, meaning that rather than relying on a futile attempt to ban or make illegal these existing models, we could instead rely on tools that determine and track their use in illicit acts.

I believe that the likely answer to our AI problem is, ironically, more AI.

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?

In general, there are few reasons why it would be prudent to differentiate between whether an entity is using an open or closed weight AI. In general, all entities utilizing AI for the processing of third party data should be required to disclose their use.

Due to privacy policies, it is generally necessary for entities to disclose their use of closed weight models because those result in the transmission of data to external vendors. Alternatively, there is likely a loophole available in which an entity may be able to avoid disclosure of open weight models, since they may be able to utilize those "in house" without the need to transmit data, and thus not need to present that information in the privacy policy.

As a blanket policy, given the delicate nature of information related to AI, all commercial use of AI in processing customer or client data should be disclosed for any entity doing so, regardless of open or closed weight. Even with open weight models – if the commercial entity hosting the models is saving the logs for the data it processes about customers or clients, all of the same issues and concerns from closed-weight models would apply. Safety, privacy, and reliability would suddenly be just as in question as a closed-weight system.

In general, I feel that any AI usage that directly affects the lives of external individuals to the host of the AI system should be reported, regardless of whether the model is open or closed weight.

An individual using AI in their home to help read white papers, or to help format their personal blog posts, would likely not meet this criteria. But a commercial or government entity that is processing data about people in order to make decisions that could affect the physical, financial and mental well-being of those individuals most certainly would.

d. What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights?

The government should absolutely be involved in the creation of best practice standards for the use of AI across the country. Similar to the International Organization for

Standardization (ISO), the US should continually be offering guidance on proper use of technologies across the country, including AI of both closed and open weight varieties.

The restriction of open weight models is already handled via licensing, and further restriction risks stifling innovation and harming consumers. I previously outlined my personal criteria for the point that I believe that the government should step in for open weight models, but I do not believe that any model of any form, closed or open weight, currently comes close to that point.

We are at a very sensitive time in the progress of this technology throughout our country, and undue burden in the form of unnecessary and heavy-handed regulation could absolutely create new monopolies within this industry, as well as set back the country as a whole in its technical race against our national competitors.

i. Should other government or non- government bodies, currently existing or not, support the government in this role? Should this vary by sector?

Currently, international standards for AI have already been defined by the International Organization for Standardization (ISO/IEC 42001:2023 https://www.iso.org/standard/81230.html). We could already consider the existence of these standards as support from other government and non-government bodies.

e. What should the role of model hosting services (e.g., HuggingFace, GitHub, etc.) be in making dual-use models with open weights more or less available? Should hosting services host models that do not meet certain safety standards? By whom should those standards be prescribed?

It would be an undue burden upon hosting services to be required to monitor, test, red team and confirm every model that is placed upon them.

Additionally, it would create an impossible burden upon individual researchers who create re-trained versions of models, such as “fine tunes” or “merges”, to attempt to meet the regulatory paperwork requirements that would need to be generated to keep safe HuggingFace or Github’s interests when hosting these models.

Any such attempts to impose these requirements for open-source software or open weight AI would stifle innovation across the industry and do irreparable harm to the United States interests in this tech space. I cannot stress enough how dangerous this path would be, not only to innovation, research and economic prosperity within this industry, but also for the

US economy as a whole. It is paramount that this country does not tap out of this tech race with foolish actions such as overregulation of internet hosting bodies.

f. Should there be different standards for government as opposed to private industry when it comes to sharing model weights of open foundation models or contracting with companies who use them?

The US government has always had different standards in the software space when it comes to safety, security and general risk management, as compared to the private sector software industry. There is no reason for AI to be any different.

In particular, it is likely important for the US Government to be exceedingly careful of its use of foreign open weight models, or any web based third party APIs for AI of any weight type.

Additionally, it is important that any government use of AI be transparent, especially in regards to law enforcement. Such use should be open and easily accessible for third party audit, avoiding relying heavily on proprietary closed-weight AI that can avoid scrutiny under the guise of Intellectual Property protection.

It is very important that any AI use that affects the safety and freedom of the American people should be scrutinized carefully to detect and mitigate inaccuracies, faulty logic and unintended model bias.

g. What should the U.S. prioritize in working with other countries on this topic, and which countries are most important to work with?

The US should prioritize making itself appealing as a provider and source of technology in this field for any country interested in utilizing emerging technologies for their own needs. As superpowers like China improve their own AI capabilities, the US may find itself with insurmountable competition from other countries.

h. What insights from other countries or other societal systems are most useful to consider?

The ISO's standards, as mentioned above, are a perfect example of what US regulatory bodies should consider creating. Overly cumbersome regulations would create a chilling effect on innovation from companies that do not have billions of dollars in capital to spend,

but an outline of voluntary standards that American entities should strive for would lay out a framework that anyone of any means could easily follow.

i. 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.

Dedicated government teams, or even a full department, for AI may be paramount to the US staying on top of this technology. This is unlike any technology we've seen before, and while Generative AI may not be the terrifying monster that the internet makes it out to be, whatever comes next could get much closer to that. The US Government is currently far too reliant on the private sector, and particularly corporations whose personal interests may conflict with the government and society as a whole, to help them understand what is going on with this. There is little reason that the state should not have an entity at the ready to monitor and understand AI at a level far deeper than it currently does.

To be frank: if you're having to ask external actors what you should do about AI, you're already off to a bad start.

j. Are there particular individuals/entities who should or should not have access to open-weight foundation models? If so, why and under what circumstances?

Any entity that has a history of egregious human rights violations should be barred from use of any AI tools. Law Enforcement across the country has been known to utilize technologies with poor understanding of their efficacy, to the detriment of innocent citizens; this has been the cause of irreparable damage to some people's lives. This, however, can be achieved via Licensing and may not require any additional legislation to achieve.

Following up on my previous answer – there should absolutely be a government entity available with the knowledge and capability of determining this for themselves.

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?

Look at how fast this technology has progressed, and in what ways it has progressed. In 2020, we never expected that Generative AI would be as far along as it is today; the leap in capability between GPT-3 and GPT-4 is beyond substantial (https://www.geeksforgeeks.org/gpt-4-vs-gpt-3/). But at the same time, the magic is starting to wane, and people are beginning to realize that Generative AI is, in some ways, not as powerful as they had once hoped it would be (https://medium.com/@multiplatform.ai/bill-gates-suggests-that-generative-ai-may-have-reached-a-plateau-despite-openais-optimism-about-386a664853dc). There is a strong possibility that Generative AI may never become as powerful as people currently fear.

In light of these extreme ups and downs that have, thus far, defied predictions of some of our most knowledgeable scholars, I do not believe that any blanket decisions made today could accurately judge the future of AI.

Rather, the best action and decision that the US government could make today is to form a committee, team or department whose knowledge of AI is guaranteed to be within the best interest of society as a whole and not just corporate shareholders. That committee should be tasked with the continual monitoring of advancements in AI technology, with yearly reports and recommendations for changes based on their findings.

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

As stated in my previous answers leading up to now, I feel that innovation should currently be our primary focus. Major corporations with billions of dollars to spend jumping through regulatory hoops would love nothing more than to see a moat created around their growing monopolies, and will likely push for regulations that impact others far more than it impacts themselves. Other countries would love to see us take a step back from this technical race by thwarting our own people's attempts at progress with costly governmental red tape, or worse yet with legislation that simply bars them from competing at all.

While there are certainly security concerns, those primarily come down to transparency and limiting the usage of AI in a situation where it is simply not capable of performing the task adequately. AI is not yet advanced enough to effectively make life or death decisions, nor should be relied upon in situations that could result in the loss of freedom of

individuals. No corporate entity should be allowed to use AI of any weight quietly in the background, processing large quantities of data about US citizens for their own personal gain, while those citizens are none-the-wiser.

Generative AI has yet to reach the point where its capabilities pose a true threat beyond what any skilled individual could do with the tools already at their disposal. Creating a governmental body whose goal is to ensure that it knows when that threshold has been reached is paramount to our long-term well-being.

Ultimately, our most important goal should be to foster growth within our economy, and to safeguard individual citizens from the misuse of AI by people who do not understand its limitations, and apply it for tasks where it can only do more harm than good.

b. 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?

Per my answer for #5: I do not believe a mathematical calculation on the computing power required to generate the model will effectively determine whether a model is dangerous or should be regulated. If we are indeed hitting a plateau in the capabilities of Generative AI, then we will begin to see diminishing returns on the computing power put into models. It could very well be that a model trained at such a high level of compute is only 20-30% more powerful than what is already in existence from the previous generation… or even the current generation.

The E.O.'s recommendation feels like a quick band-aid fix to a problem that the government does not entirely yet understand. I recommend that this time and effort should instead be focused on increasing the government's understanding of this technology.

c. Are there more robust risk metrics for foundation models with widely available weights that will stand the test of time? Should we look at models that fall outside of the dual-use foundation model definition?

Per my response to #5 – applying a more practical approach of leaderboard style testing would be far more effective than a mathematical metric on compute. Detailed testing of

foundational models, and continual monitoring of AI spaces to stay on top of emerging re-trained and merged models for detailed testing of those as well, would give far greater insight into when models would become a danger.

Imagine if, in the year 1995, we determined that any Desktop computer with greater than 1GB of Random Access Memory was far too powerful and dangerous for any individual to have. What would that have done to innovation in this country? How long would it have taken for the government to unravel that silliness, and how much damage would it have done to the United States private sector at a time when the country was in an arms race to take a leadership role in the World Wide Web?

Rather than shooting ourselves in the foot, we should take steps that will benefit the country as a whole, not just for this generation, but for future generations as well.

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?

"AI is going to take our jobs".

No matter where you look online, one of the two chief fears that resound in the hearts of people across this country are the fears of AI putting them out of work, or fraudulent AI imagery causing direct harm to themselves or their families.

In my opinion, both concerns are extremely valid. And we, as a people and as a nation, are wholly unprepared for what appears to be coming our way.

The digital art industry has already suffered extensive damage from the existence of Generative AI (https://www.businessinsider.com/ai-taking-jobs-fears-artists-say-already-happening-2023-10), and with each passing day the idea of reducing staffing by replacing workers with AI is becoming more appealing to CEOs (https://ktla.com/news/technology/1-in-4-ceos-planning-to-replace-workers-with-ai-this-year-according-to-recent-poll/).

Despite a massive strike in Hollywood not long ago over the use of Generative AI, AI companies are already pitching even more powerful products to the film industry (https://deadline.com/2024/03/openai-heading-hollywood-pitch-revolutionary-sora-1235866748/). And this has reportedly already resulted in an $800 million dollar studio expansion in Atlanta being put on hold.

Generative AI is not powerful enough or capable enough to convince corporations that it can yet replace workers at a large scale (https://www.cnbc.com/2024/03/06/generative-ai-

holds-massive-potential-but-businesses-arent-ready-yet.html), but that won't stop some from trying.

We live in a time where wages have remained stagnant for large chunks of the US population, while housing and grocery prices have soared sky high. The most vulnerable among us could be some of the first to be replaced by Generative AI, regardless of whether the tool is actually up to the task or not. Customer service representatives, call center specialists, digital artists, writers, bloggers, and many others who already struggle to put food on the table may soon find themselves with their income streams cut off as major AI corporations pitch their products as a way to reduce headcounts.

I understand, and even believe, the argument that AI will also *create* new jobs; but this leads to the question: "who will pay to train those people for the new jobs?". These are individuals who are already living paycheck to paycheck, with neither the time nor money to be able to attain training for currently existing opportunities, much less future opportunities of which we currently know little about.

By this point I'm 14,000 words into this document, so you should have an understanding that I am somewhat fond of this technology. Despite this, I must admit that we may be walking into a nightmare scenario where a large number of jobs are simply going to vanish, as if the economy were falling apart at the seams. And yet I suspect that the corporate economy will flourish and thrive in this dystopian scenario, while many individual citizens will be left with few options but to turn to the State to be able to feed their families.

We are so focused on how open-weight models can do theoretical damage to the safety and security of the nation at some unknown point in the future, while we aren't paying nearly enough attention to the damage closed-weight systems are already doing today. And our economy is going to suffer for it.

The truth is: stopping the AI sector from growing isn't an answer to this problem that would work well for us. The US is not the only country building powerful AI systems, and someone will end up at the top of this tech race. The US economy has greatly benefited from winning similar races with other technologies in the past, and I suspect that we'd like a repeat performance in this arena as well. But that doesn't mean that we can ignore the damage that AI will do if we leave its use unchecked across the private and governmental sectors.

We either need some form of regulation and oversight to stop the wholesale replacement of human headcounts with artificial workers, or we need a plan in place to cushion the fall for those people who lose their income streams to this technology. Regardless of how you feel about social welfare systems, I would struggle to imagine how you could disagree that

a sudden mass increase in unemployed and unhoused individuals might have poor results for our society as a whole.

This is the danger we need to be thinking about the most right now. And we are no longer at a point where we can ponder the problem for several more years; this problem is beginning now, and waiting until it is too late will be a massive burden upon this economy, and the people of our country.

Open-weight AI is smaller and weaker in every meaningful way to the closed-weight models that exist today. It will be a long time before our most pressing threat are these little models that can barely answer simple factual questions without hallucinating, or the open weight image generation tools that seem to believe the human hand has a variable number of fingers. But Generative AI, as a whole, and how we can prepare for its effects on our people is something we must consider carefully when thinking about the health of our nation's economy going forward.

I thank you for taking the time to read this rather lengthy and opinionated response of mine, and I hope you will consider the benefits that open-weight models will bring when deciding the fate of it within this country.

Good luck in your decision,

- SOCG