I would love for this comment to be posted, but I would like to keep myself anonymous. For context, my higher education background is in the biological sciences, and I am a hobbyist programmer. I am a citizen of the US and a resident of the state of California.

Not all of the questions have been answered. I have limited my responses to subjects that I am somewhat proficient in.

Answers to the ten questions are as follows.

1. An *open weight model* should be defined as an artificial intelligence (AI) model available for public use that can be run on personal or rented hardware independently of a third-party service. This will typically necessitate information on the architecture of the model, in addition to all weights. There may be stipulations with regards to the usage of these weights (e.g. non-commercial use), but only those models that may be distributed freely should be considered *open weight models*. This mirrors practices in open-source or source-available software: code itself may be viewed, downloaded, and compiled, but certain restrictions on its use may exist. It stands in opposition to proprietary models.

   Widely available should in theory mean weights may be freely distributed and shared, again with certain stipulations. In practice, though, weights have a tendency to be "leaked" and spread like wildfire across the public Internet, and may not be used in accordance with the model creator's wishes. Widely available and open should be synonymous.

   Will there be open models exceeding the performance of that of proprietary ones? Almost definitely, and there is historical precedent supporting this. GPT-3, from OpenAI, was first released in December 2022 and was at the time considered state of the art, drawing a considerable amount of interest and media attention. Comparable open models exceeding GPT-3's performance include Qwen1.5-72B-Chat and Mixtral-8x7b-Instruct, as measured on the LMSys Chatbot Arena Leaderboard[1]. Both have appeared within a two year timeframe. Though there are may be other benchmarks that place proprietary models ahead of these open-source ones, these benchmarks evaluate model performance on the basis of pre-written questions and answers, which may, through intention or negligence, find their way into models' training data. The Arena Leaderboard measures model performance in terms of human preference. There are downsides to this form of evaluation — the nature of the test tends to prefer short-form, direct conversations over advanced, domain-specific reasoning or recall. Absent continuous upgrades to GPT-3, it is likely that open models will continue to eclipse its capabilities.

   The question of whether models will *continue* to eclipse state-of-the-art proprietary models is uncertain, as we lack information on what the capabilities of these models are. A desire among the enthusiast community — a sort of community "white whale" for the last year, so to speak — has been an open-weight model rivaling GPT-4. Though there is no model at the present time that beats GPT-4 in all or most categories, there have been fine-tunes of smaller, less capable open models that beat GPT-4 in one category. A fine-tuned model is one that has been trained with additional data of a specific format, nature, or topic that enhances its capabilities in that domain.

2. *What if AI makes a killer virus?*
   The most common way that AI risk is described is to compare it to biological weapons.

[1] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

Imagine, they say, an AI model that can generate viral RNA sequences for a deadly biological weapon. With proprietary models, it is said, there is at least the chance that these models can be fine-tuned into refusing such requests; local models can have their so-called guardrails “eliminated” to produce such information in private.

I do not wish to be cavalier about the risks of bioterrorism, but this argument handwaves the sheer complexity of actually producing a biological weapon. I could, for instance, have the RNA sequence of a killer virus on my computer, but until I somehow produce this sequence of nucleotides in a lab, this information is harmless. *AI bioterrorism risks should be mitigated at the level of the procurement of laboratory equipment.*

Consider: you have the RNA (or cDNA) sequence of what you believe to be a killer microbe produced by a local AI model, and you are determined to unleash it on the public to further a social, religious, or economic goal. How would you:

- Get your hands on the actual RNA/DNA bases? Synthesizing genetic material *de novo* is a much more difficult process than it seems, and many labs outsource this to a company. It is these companies that should implement safeguards, such as know your customer (KYC) protocols and checks that compare requested sequences against known toxins. Likewise, access to and sale of DNA synthesis machines should be restricted.

  How is this any different from open weight AI models? Isn’t this an argument that we should regulate both open weight models and *DNA synthesis machines?* DNA synthesis produces something tangible. I can take the DNA from a synthesis machine, inject it into bacteria or even my own cells, and immediately result in some functional change. DNA is a physical, non-fungible item that I can hold in my hand and has real consequences if injected into my own cells. I cannot choose to turn off this DNA on a whim; I can turn off my AI models if I want them to stop producing information. In other words, *no open-weight AI model can directly make anyone sick — it is code on a computer.* It should be attempting to translate this information into the physical world that should come with regulations.
- Preserve and use this RNA? Because RNA is an unstable molecule, most procedures call for it to be stored at -80°C. My household freezer does not go this low; these have to be ordered from specialty vendors. Lab equipment is not cheap, and they come from specific vendors. Again, KYC. I can go to an open-weight model (or a website, for that matter — no AI needed) and ask it to tell me how to make a bomb with ammonium nitrate. I can start this process on AI, but where it should end is when the FBI gets word of a suspicious shipment of ammonium nitrate to a private residence and begins investigating.

In short, bioterrorism is a real risk and one that I seek not to minimize, but with biologicals being a *physical* entity and AI being a *digital* one, regulation should occur where the digital becomes physical.

I would also like to point out that cDNA for certain lethal biological toxins, such as ricin[2], are available online. cDNA can be synthesized, injected into cells, and expressed to form the toxin, no AI needed. Many biology majors in college perform this exact experiment — inserting cDNA into an expression vector into *E. coli* cells to produce purifiable proteins. How would you get your hands on this cDNA? How would you express it? Ask the AI all you want, but it probably won’t stop the FBI from taking an unusual interest in what you’re doing.

[2] https://www.ncbi.nlm.nih.gov/protein/XP_002534649.1

Open-weight AIs will lower barriers for entry, allowing anyone to have a personal expert assistant at their fingertips.
Before we get into this discussion, it is worth examining how these AIs are trained. Models, whether open-weight or not, are trained on a vast corpus of data derived from books, magazines, websites, scientific journals, and so on. Its knowledge of the world is derived from these sources; a model not trained on a particular book will have no information about the content of said book.

This raises the question: *what can AI teach us that we cannot learn from books and online sources?* Just as we can find the cDNA for ricin online, there are plenty of sources that instruct readers on the synthesis of illegal drugs, explosives, and toxins. They may not be easy to obtain, but anyone sufficiently motivated will be able to find recipes online. Compared to biological warfare agents produced *de novo*, the threshold to entry is much lower. Think about it this way: if I were to ask an open-weight model how to produce methamphetamine and it obliged, the first thing to ask would be how exactly the model learned these instructions. If such open-weight models were banned, I could still find this information online through a simple Google search. There is also no guarantee that these AI-produced instructions would be correct.

For complete fairness, I want to describe how this may act as a potential risk: it makes access to information easier. AI can summarize complex scientific articles into forms that are easier for the layperson to read, which these individuals may use to produce a biological weapon. I find this argument particularly fallacious because the same argument can be applied to the printing press and the Internet, yet there are no calls for widespread bans of personal printers and home computers. The Internet, after all, allows the dissemination of peer-reviewed scientific papers, some of which may focus on biowarfare. Clearly, if open weight models are a threat, the open Internet is also a threat, since it allows delivery of modern scientific literature to potentially nefarious thrat actors. Thus, by this logic, we must ban all personal Internet and allow access from only carefully monitored Internet access points.

Furthermore, I find particular benefit in giving the public free access to tools that enhance scientific understanding. Even with my background, I find certain papers to be difficult to follow and understand. Having a model capable of summarizing dense, peer-reviewed scientific text and making it accessible to a wider audience is always a benefit.

AI and privacy (Question 2C)
I can only speak for the medical field, but open-weight models, capable of being run on personal hardware, is the superior choice in terms of privacy.

Medical providers are notoriously averse to security breaches; few of them use cloud services in the fear that. In fact, a local dental office that I visit keeps their patient records and X-ray information locally on premises, and the system is secured using a lock and key. Patient privacy is incredibly important due to both HIPAA requirements and basic ethics.

I am not a licensed medical professional, but I have worked in a medical context before, and my overwhelming experience is that medical professionals take privacy extremely seriously. Few would be open to storing or processing information on an external server. Most medical professionals are not willing to store patient information unless absolutely necessary, since even with HIPAA compliance there is always the fear that they may have misconfigured some privacy setting or data breaches may expose patient data. Having open-weight models that *run* *locally* on private, secured hardware would be a boon for those expecting privacy.

This is not limited to medical settings; I find in my own interactions with my local AI models that I am more honest and cogent with them compared to remote ones, as there is always the risk that this information may be retained for future training purposes or even leaked[3], through malice or negligence. This has implications for both private individuals like myself and large corporations[4]. I can speak only for myself, but I enjoy the peace of mind that a private, on-device model brings, that my chat history will remain private, and I have every right to delete any message that I want without judgment.

I have one personal anecdote that should illustrate the importance of open-weight AI. I was in a situation in which I needed to run a bioinformatics analysis on some patient data. Writing the code myself would have taken a few hours, but with a local AI, I finished it in less than ten. I had qualms about uploading some of this data to ChatGPT or another API endpoint, even given privacy guarantees. Had I not had these open-weight models, I would simply have written the code myself, costing myself and the lab a few hours.

Couldn't open-weight AI generate large amounts of misinformation?
This is a genuine concern with the rise of AI systems. The common line of reasoning goes that with open-weight AI, individuals could generate large quantities of misinformation and spam while proprietary systems would have some safeguards. While I believe these concerns to be founded, I disagree that the correct course of action is to regulate open-weight AI models.

Firstly, any sufficiently large company desiring to target users with marketing emails will have sufficient resources to produce a model of their own. These models would be bound by existing regulations on *all* models, not just open-weight ones, but it would not be a stretch for an unscrupulous company operating outside of the US to flaunt our regulations and bombard users with spam. Regulating open-weight AI reduces the barrier to entry but does not fully eliminate it.

Secondly, I question the ability of regulation on open-weight AIs as being sufficient to inhibit spam, mainly because so much spam is already generated with proprietary AI models like ChatGPT. When ChatGPT spam is detected, it becomes big news[5]; there are countless examples of bad actors getting away with spam generated on ChatGPT.

I acknowledge the fact that open-weight models would make misinformation easier to spread and should be one possible area for regulatory concern. However, consider the fact that open-model AI systems can also assist in the detection of spam and political material. Misinformation about, for instance, climate change may become more prevalent, but advocacy groups and non-profits could fine-tune an existing open-weight AI system to combat this misinformation, summarizing recent findings in dense scientific journals in a more readable form.

Still, if the P(doom) > 0, then we should certainly discuss regulating open-weight AI systems.
With the risk of catastrophe low but not entirely zero, we should play it safe and ban all open-weight AI systems, playing Pascal's wager on AI. The problem with this argument is that

[3] https://www.livemint.com/ai/chatgpt-leaks-sensitive-conversations-ignites-privacy-concerns-heres-what-happened-11706705781882.html

[4] https://techcrunch.com/2023/05/02/samsung-bans-use-of-generative-ai-tools-like-chatgpt-after-april-internal-data-leak/

[5] https://www.washingtonpost.com/technology/2024/01/20/openai-use-policy-ai-writing-amazon-x/

applies equally to the concept of bringing the printing press, the Internet, and strong encryption to the masses, none of which have been outlawed.

- Unrestricted access to printing presses makes misinformation, libel, and hate easier to spread. While in the past, books had to be written by monks and scribes, access to the printing press can lead to the widespread dissemination of information regarding the manufacture of weapons, spam, and misleading information. Clearly, this means that the printing press must be kept proprietary, and plans and blueprints must be hidden so that the public cannot get access to them.
- The Internet makes it incredibly easy for bad actors to disseminate information online, access guides to bomb and weapons manufacturing, and communicate with like-minded people online. This means that access to the Internet must not be given to the common person, and when the Internet is accessed, we must constantly monitor for threats.
- Encryption allows terrorists to communicate with impunity, sharing plans for bioterrorism in secret chat rooms, and it also allows the undetectable transmission of illicit material, such as child pornography. Though encryption may be a boon for privacy activists, it can be deadly in the wrong hands and must be strictly regulated.

All of the above situations came with real risks and negative consequences. The printing press allowed the production of more misinformation — but it allowed the flourishing of literature and science. Disinformation on the Internet is a real and troubling phenomenon that has had numerous negative effects — but it allows for people to rapidly search for and access the near total sum of human intelligence. Encryption does allow criminals to get away with activity that they would not otherwise have gotten away with — but it allows the secure and private transmission of information between multiple parties, greatly improving our digital privacy.

Open-weight AI has many of the same risks, and I feel strongly that these concerns should be addressed. However, bans or limitations these AI systems are overly blunt and destructive, and completely eliminate the benefits that the public would reap from having access to these AI systems. A common comparison made towards open-weight AIs are that open-sourcing them would be similar to open-sourcing nuclear weapons. I disagree for two reasons:

- Nuclear weapons have one purpose: to kill hundreds of thousands or millions of people. While AI-generated bioterrorism agents *can* cause such havoc, we have (1) not seen any demonstration of any such capabilities while the destructive potential of nuclear weapons was immediately apparent and (2) the positive and productive uses of AI are far more numerous.
- Like bioterrorism, having the plans for nuclear weapons is one thing; actually building the bomb is another. Any potential terrorists will quickly run into issues procuring adequate amounts of fissile material, since production facilities are heavily guarded.

But closed models would allow the public to reap many of the benefits while eliminating many of the risks, right?

It is true that closed/proprietary models are inherently less risky than open-weight models, simply due to the additional, unremovable guardrailing. However, the question is to what degree that the risks are eliminated and how many of the benefits are maintained. As mentioned above, I disagree that the main risk comes from the AI models themselves but rather the methods in which digital plans are translated into real-world objects, since said plans can be generated without the help of AI. Large corporations are more than happy to accept customer money to generate spam and commit copyright violations[6], though these are nominally against

[6] https://techcrunch.com/2024/03/20/openais-chatbot-store-is-filling-up-with-spam/

the terms of service.

Open-weight models would allow them to produce hate speech and violent content.
It is possible to make open-weight models, especially “uncensored” models (ones with safety fine-tuning removed), generate hateful material and slurs against various groups. I am a non-white “person of color” (POC) and seeing such invectives generated against me, other POC, and other disadvantaged groups is definitely unsettling. I believe that the fight against racism and other forms of discrimination is a noble one, and AI may provide us with the tools to help combat prejudice.

I want to stress, however, that the function of large language models (LLMs) is to simply predict the next token (a word, a letter, or a portion of a word), having being trained on large quantities of text. For instance, prompting an LLM with the text “The sky is” would likely cause it to predict the word “blue,” perhaps not because of any inherent understanding of the color of the sky but because of the data that the model has been trained on. “Sky” and “blue” appear next to one another far more often than, say, “sky” and “green.” *An LLM will produce offensive content when it is prompted to do so.* An LLM is a tool. We would not want. Like with any privately owned/operated appliance, it is possible to generate offensive content. The question is whether the harm caused by this ability to produce offensive material *when specifically prompted* is greater than the harm caused by regulating these open-weight models out of existence.

Just as we do not sue the maker of the pencil when libel is penned or regulate the ability of megaphones to amplify potentially distasteful content, I fail to see why AIs should be treated any differently. The argument could be made that common appliances do come with safety features to eliminate the risk of harm, such as a table saw that automatically stops when it detects flesh. There is a difference between the two of them because (1) safety features in most appliances protect against *inadvertent* harm (e.g., sawing off a finger), whereas filters and content moderation protect against *deliberate user action*; (2) the presence of safety mechanisms in physical machines should ideally not affect the efficiency of the rest of machine, but censorship and alignment decrease performance across the board, even for unrelated tasks[7]; (3) the harm caused by table saws are immediate, physical, and universally understood, but the risk of AIs are not immediate. I cannot cause physical damage to my computer, my property, or my body solely by using an AI; I have the option of turning the model off at any time and deleting it from my computer, whereas a detached finger is permanent.

Companies should be free to release foundation models and safer/aligned fine-tunes, and fine-tune these models in accordance to their own beliefs on AI safety. The beauty of open models is that it allows users to undo or override these trainings, just as I can remove the safety mechanisms of my home table saw. If I use my Internet connection to disseminate a threat against a public official, it is me that is charged, not my Internet provider; the onus is on me to avoid disseminating information that harms other people. Individuals, not the model’s creators, should be responsible for a model’s output.

- *Why do people want open-weight AI in the first place?* (Question 3)
Having read a number of arguments against the release of open-weight AIs, I am struck by the motivations that they ascribe to advocates such as I. Though I wish to avoid painting any group

[7] https://old.reddit.com/r/MachineLearning/comments/13tqvdn/uncensored_models_finetuned_without_artificial/

with an overly broad brush, I am struck by the argument that many provide. We are not idealistic but naive ideologues that repeat the mantra “open things are always good,” to which they can respond, paternalistic and smug, to us silly children not everything that is open is good — case in point, nuclear weapons. That is true. What I also believe to be unequivocally true is that releasing these weights will provide a massive boon to the AI research, enthusiast, and entrepreneurship communities, and a world with open access to AI.

Open-weight AI is a boon to researchers.
Imagine for a second that you were running a biology experiment with little worms in a petri dish. These worms, you have measured, show remarkable intelligence, much more so than any worm you have ever seen before. The first thing any scientist would do is dissect the worms and immediately get to the bottom of this mystery. What causes the worms to be so intelligent? Is it nature, nurture, or some combination of both? In any case, you would want full, unfettered access to these worms. Now imagine if came up to you and said you could only access the worms indirectly.

This is what open-weight AI is: full, unfettered access to an advanced artificial intelligence system. The internals of many of these models are unknown. We still do not fully understand why these models are so unreasonably effective at producing text and, to a degree, even reasoning around the world. Though it is debatable whether current models can be considered artificial general intelligence (AGI), the mere fact that this term is being tossed around with any degree of seriousness is telling us that *perhaps there’s something more to these models that meet the eye.*

Having access to AI models allows their full internals to be analyzed and tweaked, not only by reputable scientists, but common citizens. The community has numerous fine-tunes, merges (combining two open-weight models together to form a larger one), and quantizations (compressing the weights in an open-weight model), all for free. Communities, such as Reddit’s /r/LocalLLaMA continuously discuss some of the more fascinating aspects of large language models, evaluating, testing, and learning. Experiments include adding noise to model weights to inspect what occurs to them[8] or modifying the operation of a model such that unexpected results or behavior is often seen[9].

Open weight AI is flexible and cost-effective.
I recently had the opportunity to mentor some students, who wanted to build an AI-powered app for a local competition. Using open-weight models, we were able to construct in a very short period of time a full-fledged app around Meta’s LLaMA 2. The model we used was small enough to fit on one of the students’ personal computers, and we chose a fine-tune that had considerable alignment and safety built into it.

Advocates of closed AI models often claim that researchers, hobbyists, and students can do with their models anything that open-weight models can do and more. After all, GPT-4 can in some ways be considered a scaled-up version of the smaller, open-weight LLaMA models that hobbyists find themselves using.

While in theory, using an API obviates the need for complex hosting setups, in practice, it leads to additional complexity. For this small, non-profit project, we would need to set up billing

[8] https://old.reddit.com/r/LocalLLaMA/comments/1b7e4mf/model_neurodegeneration_at_different_noise_levels/
[9] https://vgel.me/posts/representation-engineering/

information, API keys, and some sort of integration with an external server. The costs and complexity of setting it up far exceed the complexity of downloading a model and using open-source inference code to run it. Were open-weight models unavailable, chances are that we would not have "sucked it up" and gone with ChatGPT or another closed model; we would simply not have started this project.

The ability to use and fine-tune models however we wish is the greatest selling point of open-weight models. For the first time in history, we have a software system that can not only process natural-language instructions and reply. The possibilities are endless. The freedom to experiment with models as we wish, to be able to fine-tune our data as needed, and to deploy these models in widespread places should be a natural consequence of the development of such models. The simple fact is that some people may be reticent to use API-gated services for various reasons, and limiting access to open-weight models will not cause these individuals to shift to closed models but simply abandon the project entirely.

Banning or heavily regulating open-weight models will allow foreign actors to surpass the US.
(Question 3D)
Suppose that tomorrow, Congress passes a law banning the release and distribution of all open-weight models. What would occur?

All legal AI ventures producing open-weight AI models would likely cease. Would distributed, decentralized activities related to LLMs stop? Likely not. Production of these models would likely go underground, where regulation is much more difficult, just Prohibition pushed alcohol production underground instead of banning it, or banning the manufacture or sale of methamphetamine did not bring its usage to zero.

More likely, training of open-weight AI models will simply move offshore, out of the reach of the US's regulatory power. Though most of the West may follow along with the US, it is unlikely that the rest of the world would. Nations like China may release their own open-weight models, and these would almost have an anti-US, pro-China perspective baked into them. These would be the models that US citizens download onto their own computers. The law can ban the distribution of these models, sure, but copyrighted works are similarly protected, and there exists a vibrant world of media piracy. Reasonably speaking, the US cannot eliminate the distribution of open-weight models *once they are released* without stepping on many of our Constitutional rights. Eliminating the release of open-weight models is much more feasible because the resources to train such models exist only in the hands of a few companies, and they typically want to avoid being fined. The US cannot and should not expend considerable political capital trying to impose a worldwide ban on hypothetical AI models, when very real instances of slavery, ethnic cleansing, and war deserve our attention and concerns today.

Just as AI innovations in the US cause other countries to follow the US's footsteps, halting innovation in the US would cause other nations to quickly catch up. The rise of open-weight models has the side effect of making it difficult for competitors in nations like China and Russia to generate widespread interest in their domestic capabilities, entrenching the US's dominance in the AI space. Why would I pay to access a subpar model produced by a Russian government-aligned company when I could access a superior, open-weights AI model released by an American company? I would rather live in a world where US companies are leading the AI race.

- *Should open-weight AI go unregulated?*
  As mentioned above, I believe that there are risks involved in the distribution of open-weight AI models, and these deserve to be considered. Many of these same classes of risks are inherent in other technological developments, such as the printing press or the Internet. We as a society have determined that the societal benefits of letting ordinary people access the printing press and the Internet outweigh the risks, and yet we have attempted many times to regulate some of their excesses. Forty years ago, there were few laws concerning cybercrime, media piracy, and digital harassment; today, robust laws exist to prosecute these same crimes. Are there flaws to the Internet's architecture? Certainly: misinformation and hate speech are rife on the Internet, and our society is for the worse because of it. Yet the Internet offers so many other opportunities that it would seem absurd today to restrict access to a few, vetted institutions.

  I believe that regulation surrounding AIs should be aimed at addressing the negative consequences directly instead of trying to regulate open-weight AIs themselves in the hopes that these negative externalities do not occur. Some examples are:
  - Regulate the capability to synthesize novel nucleic acids, the sale of lab equipment, and biological or chemical precursors. The part of the AI-synthesized virus that kills you is the virus, not the AI itself. Eliminate the capability of producing viruses, and terrorism becomes moot. After all, the sequences of many dangerous viruses are already publicly accessible. The genetic sequence of the smallpox virus is available online, annotated lovingly by numerous scientists[10], no AI needed.
  - Legislate penalties on the production of deepfakes, election interference, and spam. Legislating this will be tricky in light of our nation's strong commitment to the freedom of speech, but many of these restrictions can fall under the umbrella of existing legal principles, such as slander or cybercrime. For instance, distributed denial of service attacks (DDOS) can technically be construed as free speech, but we have decided to classify them as cybercrimes instead[11]. Continuous, unwanted spam could potentially fall under this umbrella. I am not a lawyer, however.
  - Ensure that the responsibility of generating offensive or obscene material lies on the user that interfaces with computer systems, not the open-weight model itself. When a person uses a guitar to violate noise ordinances, the person, not the guitar maker, is liable. This does not mean shielding model creators from all lawsuits, but allowing courts to make a decision in times of ambiguity.

  Some regulations that I believe would be prudent with open models are:
  - Audit, but do not necessarily require making public, the training data. At the very least, model creators should ensure that illicit content, such as child pornography, does not enter the training mix. Further regulations should ensure that personally identifying information (PII) is either excluded or sufficiently deanonymized. Rather than mandate these requirements, an interesting middle ground is certification. Model creators can choose to release an open-weight model with unaudited training data, but models with, say, NIST certification would have stronger protections against lawsuits.
  - Make public information about the training and architecture of the model for the sake of transparency, for both open-weight and closed models.
  - Ensure that fine-tunes of specific models are labeled as such.
  - Ensure the enforcement of the terms and conditions of such models. Many existing open-weight models have terms that prevent them from being used to produce spam and

[10] https://www.ncbi.nlm.nih.gov/nuccore/NC_001611.1

[11] https://www.fbi.gov/contact-us/field-offices/anchorage/fbi-intensify-efforts-to-combat-illegal-ddos-attacks

misinformation. Giving agencies more power to enforce these terms may mitigate these harms while allowing legitimate uses of these models to continue doing so.

- *How can we control the spread of such models?* (Question 7)
  The simple fact of the matter is that once a model is released, the cat is out of the bag. However, labeling such releases as "irreversible proliferation" is counterproductive, since the same language is used to describe, say, North Korea's nuclear ambitions. It is true that once a model is released, it can never be *un*released. The same, however, can be said about books, ideas, news articles, and Facebook comments. There is an implicit understanding that releasing anything to the public is an irreversible act, just as releasing the information about ricin and smallpox are.

  Suppose that we had to criminalize open-weight models. We might begin with the online dissemination of open-weight models. How far would an administration be willing to go to stop the dissemination of weights? An approach similar to copyright law could be fashioned, but there also exists a thriving piracy scene. Considering that opponents of open-weight models consider the threat of releasing such models to be "extinction-level[12]," clearly we must treat it a as a munition. Unlike firearms and other weapons, however, AI models can be held discreetly on a hard drive. The only foolproof way to stop this so-called "irreversible proliferation" is to monitor every single hard drive in the world all the time — foolproof measures are required, after all, to respond to "extinction level threats."

  Simply put, people will share things they find useful, and AI models are extremely useful. I believe there to be no way to effectively eliminate the presence of local AI models once they are released.

  So does that mean that the US government should be passive observers?
  Not at all. I believe that the release of model weights is akin to free speech, but the right to free speech has limits; it does not extend, for instance, to direct threats made against public officials. The US government can and must play a role in the future of AI, but as an actor that promotes, not limits, the growth of AI, the same way our government should ideally promotes the ideal of free speech and expression.

  The US government can and should:
  - Fund research into AI safety and alignment.
  - Create an official certification that marks AI models based on their adherence to safety and alignment guidelines. This certification should be totally *optional*, but having such options would allow individuals to choose reputable models over ones that could be more risky.
  - Encourage publicly funded research institutes to produce their own models. The Argonne National Laboratory is on the catch[13]. These models can become references for future work.

  Who should host models? (Question 7E)
  Fundamentally, due to the rise of peer-to-peer distribution networks like torrents and the fact that I can send a hard drive containing model weights across the country via USPS, there is no centralized distribution network. We should allow private companies to choose for themselves how they choose to host models. GitHub, for instance, may choose only to host models that meet some government-mandated safety level. We should not force private companies to host content that they otherwise would not, and neither should we prevent them from hosting content

[12] https://time.com/6898967/ai-extinction-national-security-risks-report/
[13] https://www.hpcwire.com/2023/11/13/training-of-1-trillion-parameter-scientific-ai-begins/

that they otherwise would. Though HuggingFace is known as a place where weights can be freely uploaded, they have disabled access to certain models, such as GPT-4chan[14] (not based on the GPT-4 model).

I believe that our nation's culture of openness, entrepreneurship, and embracement of new technology have turned it into beacon of technological progress. Open-weight models have numerous risks, both known and unknown. There are also tremendous benefits. Proposals to outlaw the release of open-weight models also eliminate all of the risks along with the benefits. I believe that there is a middle ground. We can reap many of the benefits while specifically regulating to minimize the risks.

Thank you for your time. I trust that you will make the right decision, one that benefits the people of the United States and the world.

Sincerely,
A Concerned but Optimistic US Citizen.

[14] https://huggingface.co/ykilcher/gpt-4chan