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

- 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, pre-training, or deploying a model is simultaneously widely available?
- c. What, if any, risks related to privacy could result from the wide availability of model weights?
- 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?

Section summary: For the purposes of privacy and other potential threats, treat the training weights as you would the data. If the data is based on publicly-available information, there is little reason to treat the model weights as non-public.

I have worked extensively in statistical and scientific computing, although my Princeton University Press textbook from 2008 is now mostly obsolete. But as the author of a stats textbook, I feel qualified to give the textbook definition of a statistic: *a number or set of numbers summarizing a data set.* For example, the average of a list of numbers is a statistic. Or we might impose a model, like a straight line of the form $Y = mX + b$, then use textbook methods to find the slope of the line m and its offset b—a two-parameter statistic, (m, b)—that best describes a set pf (x, y) pairs. Reporting the statistic loses some of the data, but it is useful to know what the straight line looks like, because we can use it to project beyond the data set and make predictions about what Y might be given a new X. Information is lost, as the pair (m, b) might be a summary of thousands of (x, y) pairs, but one hopes the model has retained what is most useful.

Foundation model weights as per the definitions of this RFC fit the same story, despite the far larger scale. The data set is now a large chunk of The Internet, and so the parameters are in the billions. Note also that in the simple linear example, the parameters (m, b) are just two meaningless numbers until they are paired with the model in which their meaning is expressed—it is meaningless to open model weights without opening the model.

To use the formal mathematical sense of the word "information": because a statistic is a summary of extant data, *a model cannot generate new information.*

I stress this because press about AI sometimes implies otherwise, sometimes as hype. Chatbots make inferences which in a statistical model would be marked as a prediction with a given confidence of being correct or incorrect, and state them as a simple, known fact. That we laugh off incorrect inferences as "hallucinations" makes those remaining statements we know to be true from external sources look even more miraculous. Because language is generative, a model capable of constructing text can construct novel text.

To directly address (c): If a model is trained with otherwise public data, making available model weights trained on that data adds no new information.

In a paper entitled "The 2010 Census Confidentiality Protections Failed, Here's How and Why," the Chief Statistician of the US Census Bureau and his colleagues showed that millions of people could be uniquely identified by the tables published by the Census.[1] The improved model in their paper was able to pull more information from available data than researchers had been able to do before.

AI models, at least as described in ad copy, can make better inferences from data than we can as mere mortals, and even though they are terrible with the sort of tabular data provided by the Census Bureau, we might imagine that they are capable of such reidentification feats using social media posts and other text. Because of their text interface, non-statisticians may conceivably be able to do what the best and brightest at the Census Bureau could. Nonetheless, a model trained on public information, including public social media posts and other personal information, cannot create additional privacy issues beyond what a good search engine does today. The most one could say is that they facilitate the discovery of privacy issues already present.

To directly address **(a)**: Generally speaking, combining two data sets provides more inferential power than the two sets separately. To give one early and influential example, Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas at Austin, combined anonymized data about Netflix usage with data from user reviews on the Internet Movie Database to identify at least one unique individual, indicating that they could easily out their not-public sexual orientation in the process.[2] The crossing of multiple innocuous-seeming data sets to deanonymize is sometimes referred to as the "mosaic effect."

Models in the class discussed in this RFC are trained "online," meaning that a stream of new information is fed to the model, with no specified end point (except those set by logistics or arbitrary product release dates). Given a set of model weights and new data, I could continue the training process where the original team left off and continue improving the model.

Without the model weights, a "canned model", we are left with appending handling of auxiliary information to the output of an initial model. This is how users interact with generative AI systems right now: they provide a prompt, which does not modify the canned model weights, but which the model nonetheless uses to influence the final output. Prompts could include voluminous additional information, such as the entire corpus of an author. In *retrieval-augmented generation (RAG)*, a corpus is handed to a canned model, which retrieves the most relevant documents as a first step in developing a response.[3]

[1] John M. Abowd et al., NBER Working Paper 31995, `https://www.nber.org/papers/w31995`.

[2] `https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf`.

[3] The term originates with Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* `https://arxiv.org/pdf/2005.11401.pdf`, but its usage

As users, perhaps attackers seeking to deanonymize a person or otherwise cause privacy-related harms, it is unclear how much worse of a result one would get from augmenting a canned model versus retraining an open model, and we can expect that as time passes we can expect methods of augmenting canned models to improve.

To directly address **(f)**: A targeted attack finds what can be found about a certain person, while an opportunistic attack finds anybody who can be easily identified and does so. As Abowd, et al. showed, opportunistic attacks are increasingly difficult to prevent. There is no way to know what data sets the Census could be crossed against to elicit a mosaic-effect deanonymization.

It should be noted that the U.S. Census Bureau continues to conduct the Census and to report on its results. Although it is a valid goal to guarantee that no individual ever has personal information revealed via a reply to the Census, it has never been the case that such perfect, hermetic privacy has been maintained. If we required proof that it is impossible to use a data set to reidentify an individual—including by an AI bot which states all inferences with full confidence—then no data would ever be released.

In the academic world, there are methods for making data sets including private information available.[4] For example, dbGaP from the NIH maintains a repository of data sets with sensitive information, to which users can request access, after NIH review.[5] Similar to dbGaP, but more broadly available, the Census offers research data centers where approved researchers have full access to data sets with PII.[6] The Internal Revenue Service offers access to tax data for approved researchers in a very similar manner via its Joint Statistical Research Program.[7]

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

As noted in the RFC, there has been an academic movement toward reproducible research, which necessarily means open data. The motives for reproducible research apply to ML and AI as much as any other scientific endeavor. Science advances when what works can be extracted from luck, and when results

seems to be broad and go beyond what was described in the paper. Notably, the paper refers to re-training model weights after the new layer is added.

[4] See my article, *Keeping science reproducible in a world of custom code and data*, for details. `https://arstechnica.com/science/2021/11/keeping-science-reproducible-in-a-world-of-custom-code-and-data/`.

[5] `https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login`

[6] `https://www.census.gov/about/adrm/fsrdc.html`

[7] `https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program`.

are combined with other things that worked. If we want to compare the efficacy of two distinct theoretical models of the same situation, they should both be based on the same data, else we won't know whether it is the theoretical model or the data that gave better results. Robustness for a model, beginning with the simple linear model above, is typically checked by perturbing the base data and checking how much the parameters and outcomes shift, or excluding some portions of the data and re-estimating the parameters.

For this and other reasons, open data is the gold standard of reproducible research. In the case of the models discussed in this RFC, the rub is that those data sets are far too large to simply post as a ZIP file. The second-best, then, is to publish the parameters estimated from the data. The principle remains that the closer to the root of the project other researchers can get, the more they will be able to follow the scientific procedure of replicating and improving.

> - **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?
> - **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.)?

Security research is, to a great extent, an academic field. Leafing through Carnegie Mellon University's CERT database of computer security vulnerabilities turns up a great many vulnerabilities discovered by academic researchers. All security researchers, academic and otherwise, work more effectively when given the full model.

Defining security broadly as getting a computer to produce malicious outcomes, both question (b) about traditional security and question (c) about social outcomes fall under the same story. From a linear regression on, it is far more difficult to explain how and why a canned model produces unexpected or undesired outcomes, and therefore far more difficult to determine how to redirect the model to do better.

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

The primary lesson from open source software might be how wildly successful collaborative sharing, without intellectual property restrictions, has been in software and computing. Good computing methods are "invent once, use everywhere." The Internet, World Wide Web, word processors, spreadsheets, email, the basics of neural networks, almost all the software we use in the present day, originated before software was opined to be under the purview of patent law. The key results that gave us the recent AI boom, most notably the "Attention is all you need" paper,[8] are academic research effectively in the public domain.

Open licenses are incredibly effective at bringing in new talent. On Github, there are 51,400 individual accounts with a complete copy of the Linux source code, and an estimated 14,000 developers have contributed code back to the project.[9] The competition issues raised by having a small cadre of AI vendors have been well-discussed, but there are literally millions of people who have the talent and inclination to improve AI systems, and there simply is not room for them at the headquarters of four or five companies.

Government involvement in provision of public goods related to AI is appropriate and should be encouraged. Code written by government employees has been estimated to be worth, at a lower bound, \$1B,[10] and we can expect similar benefit from generate-once, use-everywhere AI models.

- **6d.** 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?

This may be a reference to how the GNU Public Licence is built around the principle that derivative works may be redistributed, but must not add additional restrictions beyond those of the GPL. Because GPL v2 and GPL v3 have slightly different sets of restrictions, in theory an author cannot combine the source code from a GPLv2 and a GPLv3 project.[11]

Among amateur hackers, for many the distinction is too subtle, others see it as too pedantic, and because linking v2 and v3 code is generally not seen as a violation of the spirit of open source, such letter-of-the-law violations are not punished. Red Hat earned its market capitalization of \$33.4B by distributing GPLv2 packages alongside GPLv3 packages.

In short, ways have been found to work around incompatibilities to push forward, and one finds that threats of litigation over fine details are best eliminated by a culture of open collaboration.

[8] `https://arxiv.org/abs/1706.03762`

[9] For comparison, Microsoft has 72,000 people in product research across all product lines.

[10] José Bayoán Santiago Calderón and Ledia Guci, "Measuring the Cost of Open Source Software Innovation on GitHub". BEA Working Paper Series WP2022-10. `https://www.bea.gov/system/files/papers/BEA-WP2022-10.pdf`. Subsequent researchers following the same method have produced higher figures.

[11] Ironically, one key motivation for revising to v3 was the USPTO's introduction of software patents.