**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, pretraining, or deploying a model is simultaneously widely available?**

Risks associated with widely available model weights are dependent on intended use case and inadvertent/intentional misuse. The superclass of risks defined in this manner can be divided into classes of risk based on type of use case and/or possible misuses, but the latter is undefinable beyond noting the scope of applications possible to extrapolate potential harms. Risks change if training data or source code is also made widely available, but such risks are linked to the data or code as sociotechnical artifacts not the foundation models themselves. Different superclasses of risks linked to each type of artifact (training data, source code, and model/weights) require both different approaches to risk and different standards for risk assessment. For example, consent for data collection/use, intellectual property concerns for code, and dual use risks for model/weights imply categorically different harms and must be assessed separately.

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

It is rarely impossible to use a general tool, such as open foundation models, to achieve a specific goal: These models always could (and will) reduce equity if doing so is a goal. It should be noted that many commercial use cases for open foundation models leverage types of inequity to, for example, better target advertisements. This issue is not unique to foundation models and thus should not be understood as caused by use of foundation models. Rather, defining desirable and undesirable (i.e., allowed and prohibited) use cases

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

Model weights from specific layers of a deep neural network have been found to capture different types of information. This information is not easily accessible at this point, but it could be feasible in the future to extract and reconstruct sensitive information from model weights. At the same time, it is well known that privacy risks related to use of foundation models are already an issue due to leakage of information from training data. This privacy risk is unrelated to model weights since it derives from prompting of the model (i.e., model behavior based on use of its weights).

**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?**

I cannot speak to open foundation models as a threat vector, but it is worth noting that, for example, large language models generate text that humans may mistake for human generated text, and that this capability points to a wide range of ways such models could support existing ways that state or non-state actors create or exacerbate security risks. In this context, actors use of these weights is no more novel than “robocalls” – the only difference is that it may be more difficult to distinguish between automatically generated dialogue and dialogue with a human.

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

Closed models pose much greater risks since their behavior (intended and unintended) is only assessable by a subset of researchers who often have financial interest in the successful use and proliferation of a model. This conflict of interest, alone, should rule out use of closed models by government agencies. Regardless of whether a conflict of interest exists, closed models are an inefficient vehicle for safe and ethical model development since they exclude a range of actors who would have an interest and the ability to perform critical evaluation of model behavior (e.g., non-profit and academic researchers studying precisely the topics/questions posed here).

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

Risks will differ to the extent that open foundation models do or do not replace existing systems and resources. If such models replace existing systems and resources, then traditional risks linked to system migration, user training, and other sociotechnical challenges associated with adoption of new technology will apply. If open foundation models are used for entirely new purposes (i.e., do not replace other types of systems and resources) then risks increase since we have no baseline understanding of expected or safe performance means. Use of developer-based metrics for evaluation of these models would inaccurately capture real world performance, even if these metrics were well-aligned with a practical goal, since variability in how humans adopt and use this technology is not well understood.

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

Bracketing concerns about national security, jurisdictional differences in model access poses the risk of generating fragmented ecosystems of model development and use. This would be highly undesirable since economies of scale are linked to shared access to the same models to, at minimum, evaluate alternative models or model performance in different contexts. That said, it is not always desirable to encourage economies of scale if risk of misuse grows as with general use of the model. This direct relationship is likely to be similar to what we observe with internet use, where even a small proportion of nefarious actors grows in absolute size as the total number of users increases.

**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?**

Privacy risks are the most likely to arise in the short term, but the most severe risks will be those associated with generative uses of open foundation models. The capacity to mimic how humans speak or write represents a potentially transformative moment for sociotechnical systems since prior facsimiles were often limited to specific contexts or could be easily discerned. These differences mean artifacts like “deepfakes” are the first step in what will become increasingly sophisticated strategies to manipulate individuals and groups in virtual environments. Strategies like these have strong potential for spillover into real world effects and I expect we will see two specific trends emerge: (1) use of these strategies to provoke individuals to harm others, leading to prosecution of these violent acts where defense attorneys will argue their client was a victim of fraud; (2) use of these strategies to provoke organizations and states into changing policies or even functional activities, leading to artificially generated policy preferences or other reforms. In both cases, risks arise from users’ actions. In this regard, access to a tool enabling these actions is not a primary cause of these risks, as any tool capable of effecting similar ends would give rise to similar risks. For example, firearms and persuasive speech both have potential for similar risks.

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

Foundation models with model weights that are widely available are beneficial to both scientific research about these models as well as their safe and ethical deployment. Generally, it is not feasible to assess closed foundation models and this violates basic scientific principles of replicability and peer review. Foundation models can only be beneficial if these models can be studied and evaluated in a scientific manner.

**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 model weights are essential for individual researchers in academic, industry, non-profit, or governmental contexts who wish to develop additional functionality. This capacity to develop functionality is what makes such models foundational, but they cannot serve this role if their behavior cannot be tested as part of the development process. Access to these models has enormous benefit for scientific research since they are designed to accelerate development while reducing costs. For example, the cost of training a model from original data far exceeds the cost of fine-tuning a model using original data: The amount of data necessary for the latter declines in proportion to the baseline performance of the foundation model for the intended use case, and a major benefit of access to a foundation model this intrinsic value as a starting point for research.

**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?**

Without the capacity to study model behavior vis-à-vis joint access to input, output, and model weights, it is impossible to assess the model’s safety, security, or trustworthiness, let alone its ability to robustly support 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.)?**

Yes, open model weights are fundamental for (1) assessing inequity embedded within a model due to extant human biases training the model to behave in a biased manner, (2) developing new functionality whose explicit goal is to advance equity.

**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?**

I am unqualified to speak to these questions, but note national security interests and human rights are too coarse of categories for assessing how diffusion of AI models will affect either. Use case is (almost) everything, and specifics will matter a great deal since possible applications are only limited by users’ imagination and technical skills.

**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?**

Having all simultaneously widely available increases transparency and enable scientific study of foundation models as sociotechnical artifacts derived from digital traces of human behavior. It is irresponsible to characterize them as more than this as benefits of a model are inseverable from its constitutive components: Lack of access to source code and training data assumes models can have benefits independent of these components, but foundation models are not simple tools and we lack sufficient knowledge/testing to treat them as such, even if simple use cases (i.e., search).

**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.**

It is vital to understand data collection and data annotation processes since both form the basis of foundation models. Lack of access to data collection processes (i.e., clear documentation of these processes, their assumptions, and their execution) prevents identification of potential structural biases in these processes. The fact no researcher is likely to intentionally introduce such biases is a strong justification for requiring related transparency and reporting. Lack of access to data annotation processes (i.e., definition of terms, documentation of protocol development, version control for annotation protocols, justification for selection of annotators, annotator characteristics and working conditions) prevents assessment of the ethical provenance and scientific validity of the data generating processes used to train foundation models. For example, a lack of annotator diversity is a known problem for reducing bias in AI models, and it is impossible to determine if this could be a source of model bias without understanding who annotated the data, how they were selected, why, and so on. At a fundamental level, transparency around data annotation is a net benefit for foundation model developers (and others) seeking to understand and/or correct biases within these models.

**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?**

I am not qualified to assess these similarities or differences, but as a user and researcher I see no fundamental distinction and believe lessons learned from open-source software development are likely applicable.

**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?**

Availability of model weights enables market-based development of AI-based products. Without wide availability of model weights, a cartel of private corporations with sufficient capital will be in the position of determining which products are developed (and by whom) through criteria they set based on private goals which may run counter to the public benefit of a competitive market for such products. There should be no “kingmakers” in AI since it is fundamentally a technology based on linear algebra and allowing those with the resources to perform significantly more linear algebra operations than those without these resources (i.e., individuals, universities, and most private businesses) creates an artificial barrier to innovation. Such gatekeeping around how fundamental mathematical principles are applied in real world use cases runs counter to how we understand competition as an economic motivator.

**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?**

Intellectual property-related issues influence competition, benefits, and risks to the extent that a foundation model developer can (1) shirk responsibility for poor model performance causing a downstream harm due to use in a sociotechnical system developed by a 3rd party, (2) litigate use of the model for reasons the developer did not intend, and/or (3) limit access based on financial motivations. Shirking is my greatest concern, as it shifts the legal burden of responsible model development from foundation model developers to developers who, almost definitionally, have less resources to evaluate foundation model behavior to ensure safe and ethical deployment. I strongly believe public claims to have developed a foundation model entail legal commitments to ensuring its safe and ethical use, otherwise it is unclear how foundation model developers have incentive to devote resources to this task. Ironically, it is likely the organizations developing such models are uniquely positioned to have (or have the ability to obtain) the resources needed for this task.

**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?**

I see no reason these models constitute intellectual property different from other complex yet general purpose tools already in use, such as CRIPSR.

**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?**

**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?**

Easily enforceable regulatory mechanisms around licensed uses based on public/governmental criteria for allowed and prohibited use cases.

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

Wide availability of open foundation model weights will facilitate government action in AI regulation since independent developers have strong incentives to study and identify weaknesses in these models. Doing so is a signal of technical prowess and thus economically valuable to many such individuals. A positive, unintended consequence of such incentives is leveraging of market mechanisms (e.g., demand for developers who can work with these models) to support identification of issues salient to AI regulation without the need for government intervention or related funding of similar research/evaluation.

**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?**

Use of open foundation models should always be disclosed, and the availability of weights is an important part of this disclosure: If weights are not available then this signals a specific kind of evaluation has been performed (i.e., closed, limited number of individuals involved, potential financial incentives, etc.).

**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 U.S. government should create a NIST-like organization to (1) study what metrics for risk are appropriate for different use cases, (2) collect, organize, and make available an exhaustive list of use cases, (3) require reporting of appropriate metrics for each use case by foundation model developers, and (4) publicize this information for public consumption. I strongly believe the U.S. government should protect the rights of individuals and researchers to access model weights given strong potential for foundation model developers to restrict access for reasons unrelated to ethical, moral, technical, or legal considerations. A new NIST-like organization is needed for this purpose since this ability to protect the rights of individuals and researchers will be directly linked to identifying/defining ethical and legal standards for AI development by both individuals and organizations.

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

The Office for Human Research Protections should be a primary point of support for the government in this role, regardless of sector, whenever data involving human subjects is involved (e.g., data generated by humans, data generated about humans, etc.).

**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?**

Standards should be prescribed by the U.S. government as has been done in the European Union.

**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?**

No differences should exist except insofar as (1) government data has unique risks associated with its release and (2) private industry requires maintenance of viable, profitable strategies for model development. Both can be addressed appropriately using existing knowledge about risks to individuals in both contexts.

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

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

**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.**

The Common Rule should apply to all research involving foundation models, regardless of whether this work is government funded or not.

**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?**

Those who have been convicted of fraudulent use or misuse of open-weight foundation models should not have access to these models. This includes intentional and unintentional harms from model use if specific standards for model testing and safety are not met. These standards must be created and should include exceptions for unintentional harms below a threshold of financial or other damages.