Holistic AI's Response to the National Telecommunications and Information Administration's (NTIA) Request for Comments on Open-Weight AI Models

26 March 2024

National Telecommunications and Information Administration
1401 Constitution Ave.
NW Washington, DC
20230

RE: Holistic AI's Response to the NTIA's Request for Comments on the potential risks, benefits, other implications, and appropriate policy and regulatory approaches to Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

1. Holistic AI is an AI Governance platform with a mission to empower enterprises to adopt and scale AI with confidence. We are a multidisciplinary team of AI and machine learning engineers, data scientists, ethicists, business psychologists, and legal and policy experts.
2. We have deep practical experience in auditing AI systems, having assured over 100 enterprise AI projects covering more than 20,000 different algorithms. Our clients and partners include Fortune 500 corporations, SMEs, governments, and regulators. We work with several organizations to conduct independent evaluations of AI systems through our proprietary Software as a Service (SaaS) platform for AI Governance, with solutions for Risk Management, Audits and Regulatory Compliance.
3. We welcome the NTIA's Request for Comments on the Risks, Benefits and Policy Implications of Artificial Intelligence Systems with Widely Open Model Weights. We are dedicated to assisting the NTIA in achieving its objectives of gaining deeper insights into the ever-evolving discourse on AI Safety and Model Release by providing evidence-based insights and recommendations to influence better policy outcomes ahead.

2. Comments

This document provides responses to the Request for Comments issued by the National Telecommunications and Information Administration (NTIA) on the policy implications of foundation models with widely open model weights. As the NTIA actions its mandate pursuant to Executive Order 14110 in preparing a report for the President on the potential benefits, risks, and appropriate governance approaches for such models, its policy recommendations should be cognizant of the many nuances associated with managing the many risk-benefit trade-offs associated with responsibly releasing model components, as well as how they manifest at different points of access.

As such, in our responses we survey the discourse around open weights and model release by providing insights grounded in academic research, as well as from practical industry experience on the benefits and risks of making model components available at different levels of access, as well as their related broader socio-technical considerations. With our expertise in algorithm auditing and AI risk management, we largely premise our responses on the fact that the release scope of powerful foundation models with open weights should be intricately linked to safety calibrations, actualized through a proportionate blend of internal and external evaluations.

Corresponding to the different questions posed, we provide relevant recommendations for the NTIA's consideration. Furthermore, while each of the questions in the NITA's request for comments comprises multiple subparts, in our responses, we answer these implicitly rather than addressing each element for brevity.

2a. Question 1: How should the NTIA define "open" or "widely available" when thinking about foundation models and model weights?

What constitutes an open foundation model has long been a topic of debate, with disagreements ranging from the definition of 'openness', its conflation with open-source models, to its implications on model safety and downstream innovation. As such, there is currently no widely agreed upon definition of what constitutes an open AI system. In theory however, an open system in principle may be transparent, reusable, portable and inspectable, but how this is achieved in practice presents significant differences.

It is usually easier to first define what an open model is not: a fully closed model whose entire system is inaccessible outside the AI developer organization. Historically, the main commonality among the most capable models was that they were all closed, but to differing degrees. Now, most are somewhat accessible but typically through paid APIs.

The large datasets and compute needed to train the model are typically only reachable by closed, commercial labs with huge financial resources. The emerging nature of foundation model research also meant there is a limited amount of AI practitioners available to develop these models, and many of them are hired by the large, well-funded labs. Such conditions mean it is incredibly difficult for model development to be done in typically low-resource open-source environments.

However, as the landscape of foundation models rapidly develops, *who develops these models, and how that is done* has become increasingly more dynamic**.** There is now widespread acknowledgement that this technology will bear great impact (many of which are beneficial) across societal domains, and as a result, decisions around its future will be incredibly impactful. The concentration of powerful foundation models within a handful of companies has prompted a growing push for more people to have a say in the governance of the technology. In the wake of the 2018 'techlash,' the idea of openness in foundation models refers as much to the market ecosystem in which they're developed, as it does to their accessibility to the public.

It is within this context that public discourse has both accurately and inaccurately appropriated the ideas of open-source software. Open-source software is defined as software for which "the human-readable source code is available for use, study, reuse, modification, enhancement, and

redistribution by the users of that software." Indeed, many so-called open foundation models such as BLOOM language model from Big Science qualify as open source based on this definition. However, the ambiguity of the openness debate has seen many foundation models inappropriately claim open-source status because a certain component is open, but their source code is not. For example, open weights may enable broad application development but if the source code is closed, it does not enable broad reusability. In short, the difference stems from whether opening access to, or releasing a certain model component has a *similar impact* to open sourcing in that it may enable certain principles (reusability, portability, etc.). Nonetheless, given that foundation models are comprised of a variety of different components, placing them on the same binary as the open source versus proprietary debate is somewhat misleading. Thus, categorizing the openness of foundation models based on a gradient, as first proposed by Solaiman (2023), accounts for the many nuances concerning the issue.

As depicted in **Table 1,** AI providers may release any number of elements to the public, including access to the model itself (which includes the model weights), components that enable further risk analysis, and components that enable model replication. This demonstrates the complexity of the openness of models and delineates with the straightforward nature of open-source software releases.

Parts of an AI system in a release	Elements included
The model itself	• Model weights • The ability to query, adapt, or otherwise examine and conduct further research into the model
Components for Risk Analysis	• System risks, training data, fine-tuning data, and information on people involved in adapting the model through human feedback • Evaluation results from tests that researchers may have run on the base model
Components for Replication	• Technical paper detailing the training process and code

Table 1. *AI system releases, and the elements included in each release.*

The combinations of different elements released can be categorized into six levels of access to a system: fully closed, gradual/staged release, hosted access, cloud-based/API access, downloadable, and fully open. Each stage differs by the degree to which users can access a system (the components available), for what purposes, and when they are released. Therefore, access can refer to *who* can access the system regarding users; for example, a fully open model may only be accessed by vetted researchers. Access can also refer to *which components* are released; if only the model's weights are open, but the rest of the components are closed, then access to the system itself is still somewhat restricted.

Once the models are fully open and downloadable, developers cannot rescind or change the model's release. The many different components of a model, and the degree its accessibility changes, show that its 'openness' should not be viewed as a singular concept centered around one variable (such as model components), but a combination of facets: model components, and the extent and method of user access.

Recommendation: Within this gradient, the NTIA should be cognizant of the distinction between a model's openness and its accessibility. The ability for a user to download a model or any of its components plays an essential role in how widely available it is. Once this is possible, a user can run the model on their own hardware and the scope of procedural use controls put into place by the AI provider is significantly reduced. This results in a higher risk of malicious use and compliance failure because a developer organization or any other evaluator cannot truly monitor how these models are later employed. However, if the AI developer adequately audits and evaluates its model prior to opening it or any of its components, the risk of malicious use may reduce.

As an organization committed to operationalizing the responsible adoption of AI, we view some degree of openness as an enabler for audibility. Openness often improves auditability because the training data and the codebase to reproduce models can be more closely scrutinized. However, we would emphasize that unlimited downstream access – particularly without adequate prior evaluation – may interfere with safe model deployment at scale. Therefore, it is crucial that foundation models undergo thorough auditing and evaluation interventions as they pass from one end of the gradient of release, to the other.

Holistic AI also contends that the idea of 'openness' should extend to the process that determines whether a model or its components should be opened. The pathway of a large model's release is typically decided behind closed doors, with little information released to the public on how or why a provider reached a particular decision. If audits and evaluations are implemented at each stage of a release, then the provider would at least hold documentation on its decision-making process, should it decide to communicate that with a wider audience. Furthermore, this form of openness should include a wider range of participants involved in audits and evaluations. In short, these processes should not only be done only by a closed team within the AI provider, but with a concert of external evaluators to stress-test and validate the same. Model behavior should reflect diverse and multidisciplinary perspectives, potentially including the views of AI actors outside the organization. This is not to simply uphold the idea of openness, but to reduce the risk of unintended harm, as noted by the National Institute of Standards and Technology (NIST)'s AI Risk Management Framework (AI RMF).

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

Open weights are essentially model parameters and weights that are made freely available with minimal or no restrictions on their access, modification, or redistribution, to be used in a wide range of applications. In contrast, closed or restricted weights (or those with gated access) are those that are kept proprietary, or released under specific conditions of use, redistribution, or modification. There are relative benefits and risks to both these configurations: Making model

weights publicly open allows for more collaboration in AI development, distributes control to empower smaller groups and represent wider interests, and democratizes AI more broadly to ultimately drive innovation. The last point is attributable to the fact that weights can be leveraged to customize a wide range of downstream applications. This can be done through a range of adaptations such as quantisation (a technique to reduce the memory requirements and computational costs of using neural networks), finetuning (the process of adjusting a pre-trained model's parameters to optimize its performance to new use cases), and pruning (the elimination of unused parameters from a model to improve a its speed and, at times, other metrics like accuracy).

However, there has been increasing evidence in favor of gating access to model weights. In addition to helping protect a provider's intellectual property and facilitate conformance with ethical or legal standards, there are safety risks associated with open model weights. Highly capable models, for instance, have tended towards capability overhang – demonstrating new, unpredictable, and unintended capabilities with advanced progress which poses significant challenges for a model's downstream safety implications. Further there is a risk of downstream users training models to (inadvertently or maliciously) demonstrate dangerous capabilities such as autonomous replication, deceptive alignment, self-reasoning, self-proliferation and cybersecurity risks, which can have profound economic, political, and societal ramifications. Additionally, Large Language Models (LLMs) have been found to 'drift', meaning that their performance and behavior may vary over a short period of time, further indicating unforeseen changes in model behavior.

That said, increasing access and openness can also help ameliorate model risk management outcomes, as it helps enable external oversight and evaluation to hold developers to account, as well as create better performing and safer models. Such access also facilitates community-driven research, testing and investigation which in turn makes external evaluations more robust. This is particularly true since current decisions about auditing AI systems are made by individual developers, who can ultimately choose known or 'friendly' auditors. Indeed, such interventions are useful as they tap into the wider AI community to identify issues that might otherwise go unnoticed. For example, model weights are crucial to understanding interpretability, security, and safety, and transparency can facilitate deeper investigations and scrutiny of models for more comprehensive risk evaluations.

While making model weights available can create these benefits, it has also been argued that there are other, safer methods for achieving similar outcomes, which we detail in *section 2e*. As releasing model weights is an irreversible endeavor - the aperture of potential harm increases significantly, with malicious users able to share weights through peer-to-peer distribution, effectively negating any restrictions or safety measures that a developer may impose after release. Researchers also recently discovered that access to model weights can help adversarial actors jailbreak a system by making the model obey commands that result in harmful outputs. This method was shown to be transferable across LLMs such as LLaMA-2, GPT-4, Bard, and Claude, signifying that models possess similar vulnerabilities which can be easily exposed. Additionally, when weights are released along with other model components such as training data or source code, the potential of risks further increases as users have even more access to a

system with fewer restrictions. For instance, without any restrictions, models can be fine-tuned specifically for malicious uses as safety filters built into the inference code can be easily deleted or circumvented. Such targeted finetuning is also much less expensive and requires lesser compute than building an entire model, meaning that accessibility to resources might not be as big a barrier for malicious actors as previously thought. as previously thought.

Recommendation: Given the above and acknowledging the careful trade-off to be made on the risks and benefits of releasing model weights (as well as other model components), we recommend that standardized socio-technical risk assessment procedures for auditing and evaluating AI systems be established to ensure best practices for model governance and oversight. Detailed in section 2e of our response, these risk assessment procedures must be followed at every stage of release to guarantee the security of the model and allow developers the opportunity to implement any safeguards they may have overlooked, as well as test for the possible impacts of release.

2c. Question 3: What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?

The conversation around open and closed foundation models is not a dichotomy between the two, but rather a spectrum along a gradient of their release. Foundational work by Solaiman (2023) on the levels of access to generative AI systems lays the groundwork for a discussion on the benefits and risks of the openness of a model, and the way in which a model is released directly shapes its eventual societal impact. A fully closed model only permits the developer or developing organization access to the model. This protects creator's rights and intellectual property, and controls applications of the model. The benefits to opening a model, however, are manifold. This allows for a democratization of the AI ecosystem – distributing decision-making power among more voices and expanding the range of actors who define what system behavior is considered acceptable. Opening models (or their components) also drives innovation and competition, reducing market concentration. More access, customizability, and local inference expand how models are used to develop applications. The potential of local inference and adaptability also allows downstream developers to adapt and finetune models more easily and without privacy or data protection concerns as they can access model components locally. Increased access largely allows for more transparency, accountability, and reproducibility, enhancing monitoring mechanisms (especially third-party monitoring and auditing) and accelerating research. The risks and benefits of releasing models with a wider range of components are explored more in-depth in section 2d.

Figure 1 below illustrates the benefit mapping of fully closed models and models with widely available weights, where probability refers to the likelihood of each benefit being achieved within the respective model type and impact refers to the effect that each benefit has on wider society if a model remains closed or its weights are released. This diagram allows for a visualization of the benefits associated with fully closed models or models with weights available, which can help inform the NTIA's policy and regulatory recommendations, as well as help developers make appropriate decisions on model release.

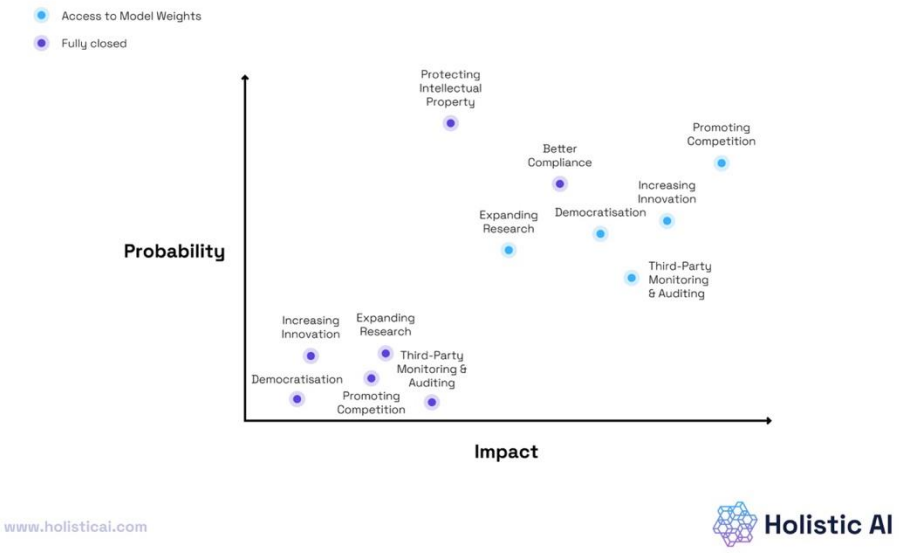

Figure 1. *The benefits of keeping models closed versus making model weights open and the associated probabilities of these benefits being realized*

2d. Question 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.

In order to precisely assess the risks posed by foundation models, it is important to analyze risks against the gradient of release to determine where they might arise, and how. Based on our previous responses on the risks and benefits associated with publicly available model weights, we emphasize that risks depend both on ***how models are released,*** as well as ***which components are made accessible***.

As illustrated by Brammer (2023), most model risks increase with progressive openness. The common risks associated with increased levels of access include **malicious use**, which can be a range of harms to political, physical, and digital systems; **capability overhang**, which is when models develop unforeseen capabilities and create further potential to drive other risks and harms; **failure of compliance**, where controls set by the developer such as verification mechanisms or other technical or legal guardrails become hard to enforce downstream; **humans being taken out of the loop** as models become increasingly autonomous and act without human verification (for example UC Berkeley's Gorilla LLM81, a fine-tuned LLaMA-based model which can call other APIs without human oversight); and **reinforcing bias** where models trained on biased data continue to reinforce said biases. Overall, these risks tend to increase as levels of access to a model increase. Only the risk of reinforcing bias fluctuates with increasing access to a model. This is because even if developers release subsequent models with mitigated levels of bias, it is not guaranteed that users will stop using the original biased model and there is no way for developers to identify and encourage users to use updated models.

As the level of access to a system increases, usually, so too does the amount of system components released, and the subsequent ease of modifying and customizing a model. For

instance, when weights are released with model architecture, almost anyone will be able to use a model that has been pre-trained to perform tasks even without access to inference code. Inference code can be easily written by LLMs and does not have to be identical to the original inference code, making it easy for actors to modify and reproduce the model. Generally, it is the process of training a model that requires the largest amount of compute and data that is likely only currently possessed by big tech companies and states; hence, the capabilities required by downstream actors to perform model adaptations are lesser if pre-trained models or model components (such as weights) are made widely available. However, even if weights are not released, it is still possible that external deployers may discover a set of weights different from those used by the original developers and can use these weights to run a model that may end up performing just as well or even better than the original one. It is thus imperative that further research is conducted to effectively pinpoint and understand what actions can be taken by external actors with varied access to different components, and the risks that can arise at each level.

As advanced by Kapoor et al. (2024), nuances in the debate of the risks surrounding open models can also be seen through the exploration of marginal risk, defined as the extent to which intentional misuse of open foundation models exacerbates risks 'relative to pre-existing technologies, closed models, or other relevant reference points.' The study found that overall, insufficient evidence exists to meaningfully characterize the marginal risks of open models to existing technology and other relevant reference points. However, for certain types of model misuse such as digitally altered non-consensual intimate imagery, open models can pose considerable marginal risks. Alongside marginal risks, the marginal benefits of openly releasing models must also be analyzed to get a more well-rounded picture of the landscape.

Another aspect of the debate is whether the same benefits (as discussed in our previous answer) can only be achieved through opening models, or whether there are other methods that could lead to similar results. Foundation model release has been likened to open sourcing in the traditional software sense, which incorporates defensive actions to guard against misuse through licenses that specify how content can be used, modified, and shared. While the common argument for open sourcing is that it allows developers to identify vulnerabilities and implement safeguards before malicious actors do, the same might not apply for AI models, especially larger and highly capable ones. There are thus contentions about the term *open source* being applied to AI systems and proposals for different types of licensing frameworks, such as Open & Responsible AI licenses (OpenRAIL). This body of research finds that the risks of open-sourcing could outweigh its benefits, especially for highly capable models that are emerging, and there are less risky methods for pursuing the same benefits.

Figure 2 illustrates the risk mapping of models with gated and non-gated downloadable access to represent the various risks presented by models with widely available weights and other relevant components. Probability refers to the likelihood of each risk proliferating within the respective model type and impact refers to the effect that each risk has on wider society for each model type.

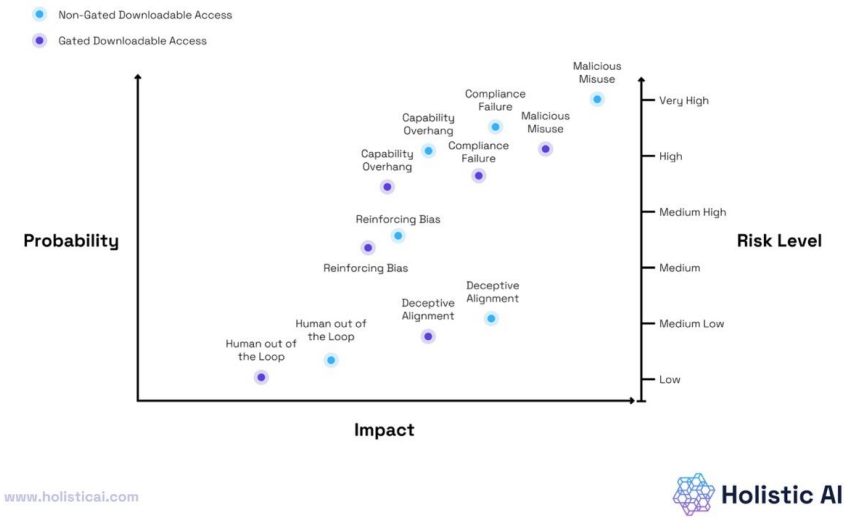

Figure 2. *Risk Mapping of Models with Weights: Gated and Non-Gated Downloadable Access*

Recommendation: We contend that it would be beneficial for the NTIA to build a risk register based on the above analysis and risk mapping diagram to help organizations identify the risks associated with their model releases and make decisions accordingly. Furthermore, given that research in this topic is still nascent, and there is a need to better understand the series of trade-offs that must be made between marginal risks and marginal benefits of releasing a model, we recommend that further analysis should be done on this subject. As such we provide two more forms of marginal risk determinations for the NTIA's consideration in the next section.

2e. Question 5: What are the safety-related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely open model weights?

and

Question 8: In the face of continually changing technology, and given unforeseen risks and benefits, how can governments, companies, individuals make decisions or plans today about open foundation models that will be useful in the future?

As mentioned in previous responses, foundation models with open weights comprise a spectrum in themselves, with varying levels of openness along the gradient of release. In line with Bommasani et al. (2023) and Solaiman (2023), these can range from having only weights available, to weights, data and source code available with usage restrictions (gated access), to components released without any restrictions (non-gated access), as can be seen in Figure 3 below.

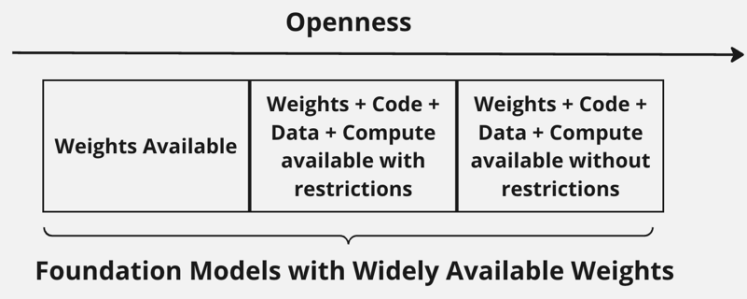

Figure 3. *The spectrum of openness of foundation models with model weights (Adapted from Bommasani et al. (2023).*

As strategies for model release progress along this spectrum (as well as the larger gradient of release) towards complete open-sourcing, providers of such models face a series of trade-offs between the benefits and risks involved. This process of determination is a complex undertaking, as there is no one universal approach when it comes to addressing them. Furthermore, there is a noticeable lack of robust and sufficient evidence on the risk profile of models in different stages of the gradient landscape, how these compare to risks generated from closed models (or even social media platforms), and how these risks balance with the benefits that come with progressive openness – adding further uncertainty, ambiguity and variables to this exercise.

Navigating Trade-Offs along the Gradient of Release: Determining Marginal Risks

To effectively interrogate and embed these considerations, it is crucial for policy and governance discourses on responsible model release to be anchored around the concept of the marginal risk posed by open foundation models. These can be relative to that posed by closed models and pre-existing applications like social media platforms, and along the evidence gaps mentioned, should also be extended to the extent of a model's relativeness openness along the gradient of release, as well as relative to the corresponding benefits posed by such models. As such, determining the extent of a model's release (both along the gradient as well within the spectrum of models with open weights) based on such marginal risk determinations can provide a credible pathway to validate tolerable levels of openness. Applied to the sub-spectrum of open foundation models, this can also help developers navigate trade-offs in opening different components associated with model weights and parameters. Parallelly, it may be worthwhile to develop similar approaches to determine the *marginal benefit* of open foundation models along the gradient of release.

Recommendation: An effective combination of these determinations can be useful in helping model developers and providers articulate the **explicit purpose** of exposing different model components to the external world. As such, we recommend that the NTIA actively solicit the input of experts to collaboratively develop such typologies.

Applying Gradient-Based Safety Strategies

Besides effectively managing risk-benefit trade-offs, articulating the explicit purpose of openness, and providing robust justifications for access based on such analyses, we also

contend that this approach can aid in determining proportionate and technically feasible risk mitigations and investigations based on the gradient of release. Not only does this help spotlight appropriate safety strategies to guide responsible release in a way that mitigates potential harm and promotes open-source innovation, it also interrogates (and by doing so, complements) the notion that a high degree of openness is a crucial precondition to validate a model's safety.

As highlighted by Seger et al. (2023), while there are safety benefits in open-sourcing models (through increased community-driven evaluation, richer auditability and the identification of unknown risks through adversarial testing), the offence-defense balance for the same is likely to skew towards offence for highly capable foundation models. Components may also be restricted due to concerns around intellectual property (IP) rights, privacy and consent. Additionally, while increased openness can aid in addressing certain complex safety concerns like capability overhang and deceptive alignment, it may not be as effective for tackling discrete issues such as interface glitches, authentication problems, and self-contained flaws. Given the substantial resources required for such endeavors, this points to the acute need to deploy proportionate safety strategies that may, or may not necessitate enhanced openness. **A thoughtful blend of access-dependent and independent approaches is therefore warranted to gauge the safety levels of a model, guiding its responsible release along the gradient.** Additionally, while internal safeguards deployed by model providers have been successful in operationalizing model safety and risk management, it is crucial to embed external oversight and validation mechanisms. Audits are pivotal in this context, as they facilitate robust system assurance by conducting impartial and independent evaluations.
We illustrate this concept in the Figure 4 based on the versions of the gradient of responsible release advanced by Solaiman (2023), Kapoor et al. (2024) and Brammer (2023). Further, given the irreversible nature of model weights release, we provide insights on how these safety strategies, triangulated with external auditing mechanisms and regulatory guardrails can help reasonably mitigate risks, confidently helping model providers release model components and increase access over time.

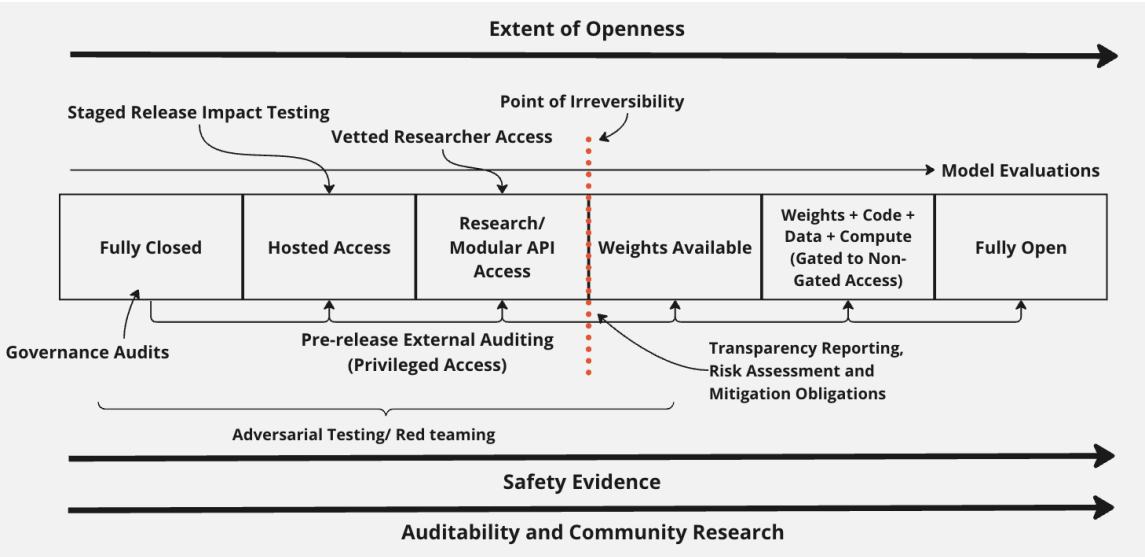

Figure 4. *Proportionate safety strategies along the gradient of release.*

Recommendation: In line with Shevlane et al. (2023), we contend that decisions to increase access and/or release model components should be anchored in the level of safety evidence gathered through a mix of technically feasible, proportionate and gradient-based evaluations, validated through independent external audits – wherein as a model's safety evidence increases, so should its exposure to the external world. To this end, we recommend the inclusion of the following measures:

- **Continuous External governance audits** of the processes, organizational systems, oversight mechanisms and quality management systems involved in model training and development, benchmarked to standards (e.g. ISO/IEC 42001, NIST AI RMF)
- **Model evaluations** through internal, academic and aggregate benchmarking to ensure a model is achieving baseline levels in terms of its safety mitigation capabilities
- **Staged-release Impact Testing** (which does not require the release of model weights) at hosted access levels, wherein developers can release progressively larger iterations of a model for safety testing through APIs to gather observational data about a model is likely to be misused and modified if open-sourced, with requisite release periods stipulated to ensure adequate baselines of research and investigation is being conducted
- **Institutionalizing Research and Modular APIs, or gated downloadable access** for vetted researchers to test a model's risk management capabilities in instances where access to weights is required to experiment with fine-tuning
- **Privileged model access** to trusted auditors to test verify and validate a model's safety and security levels prior to a model component's (or a set of components) release
- **Systematic Adversarial Testing/ Red Teaming** exercises leveraging external practitioners with multidisciplinary expertise to stress-test models for deceptive capabilities, security loopholes and potential vectors of misuse.

Determining the release of Model Weights

The release of model weights is irreversible, with providers effectively relinquishing control over its subsequent deployment. Malicious actors may leverage this irreversibility and may seek to abuse and modify model weights to further dangerous capabilities. Given this risk, we find it crucial for policy mechanisms to stipulate the fulfilment of a credible baseline-level of sociotechnical safety testing, external researcher access, external auditing and the results from these interventions as crucial determinants of whether model weights should indeed be released. These could in turn help in establishing *Responsible Release Policies* (similar to the concept Responsible Scaling Policies developed by METR (formerly known as the Alignment Research Centre (ARC)) - that could provide thresholds of acceptable release levels linked to a developer's current protective measures, as well as indicate zones of dangerous capability where model release should be paused until appropriate security and safety measures are developed, improved and productionized.

In addition to this, regulatory guardrails in the form of deployment risk assessments, transparency, and audit reports should also be required of developers of the *most capable foundation models*. In terms of deciding what these guardrails are – while compute thresholds

measured in Floating-Point Operations (FLOPS) provided in the Executive Order 14110 (10^26) and the EU's AI Act (10^25) may be starting points to determine their inclusion criteria and consequent regulatory burden, they may not be adequate as sole determinants – given that smaller models trained with fewer FLOPs (e.g. 10^24) may be equally susceptible to safety and security risks.

Recommendation: We recommend that the NTIA collaborates with NIST to co-develop a thoughtful combination of determining factors (which may include insights gleaned from Audit reports, Deployment Risk Assessments, Daily/Monthly Average Users as well as appropriate compute thresholds) to qualify models as such.

Essentially, the release scope of powerful foundation models with open weights should be intricately linked to their safety calibrations, and bodies like the NTIA and NIST are well-positioned to provide guidance in this regard. We believe that leveraging resources such as the forthcoming NIST AI 100-1, as well as multistakeholder expert coalitions such as the US AI Safety Institute Consortium, may serve as valuable agents for guiding responsible development, release and deployment, while also contributing to grounding the open-closed source AI debate on safety calibration, evaluations, auditing and risk management. We provide an initial framework to help guide future research efforts on the same below.

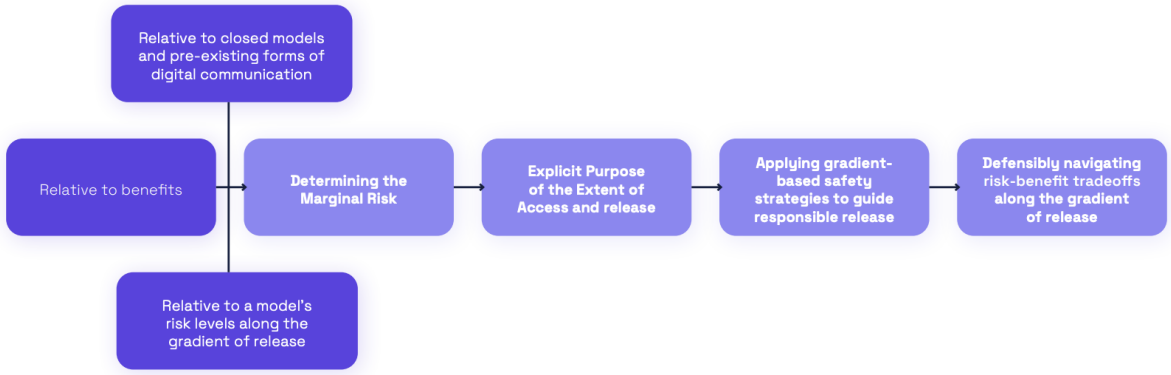

Figure 5: *An initial framework to navigate risk-benefit trade-offs along the gradient of release.*

2f. Question 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 entities should take a leadership role across which features of governance?

We list some of the key regulatory developments and mechanisms on foundation models with available weights below. Our response largely focuses on voluntary mechanisms in the United States, the European Union (EU) AI Act, and self-regulatory initiatives.

The NIST AI RMF and Executive Order 14110

The National Institute of Standard and Technology (NIST)'s AI Risk Management Framework (AI RMF) is critical as it can be applied across foundation models with varying degrees of

openness. Furthermore, its structured yet flexible approach accommodates for the multitude of users and organizations, purposes, and capacities that will deploy foundation models. As experts in AI Governance and Auditing, we appreciate that the AI RMF emphasizes auditing practices. The RMF Playbook's Suggested Actions recommend that organizations establish frequency and detail for auditing and review processes as well as the public disclosure of such processes. Foundation models with widely available weights, or those that sit farther along the gradient of release are well-suited to these recommendations. Such models provide a higher degree of scrutiny and transparency, enabling a straightforward implementation of the AI RMF.

Both the NIST AI RMF and Executive Order 14110 in general acknowledge that a variety of downstream players will incorporate both open and closed foundation models into their products and services. As such, Section 8 of the Executive Order recognizes that compliance policies should take on a sector specific approach. It instructs the appropriate federal agencies to enforce 'technology agnostic authorities' to minimize harm to consumers. Holistic AI appreciates this approach as we work closely with enterprises from a variety of different sectors and therefore recognize the nuances in risks that appear across domains.

The EU AI Act

The EU AI Act's provision on open-source models will have a significant impact on the open foundation model ecosystem. Under the Act's latest iteration – which received parliamentary approval on March 13, 2024, foundation models have been categorized as General Purpose AI (GPAI) models. Here, GPAI models refer to core model components that, while central, do not comprise complete AI systems. It is only when these models are combined with other elements, such as user interfaces or feedback mechanisms, that they form a fully functioning GPAI System (GPAIS). Therefore, when a GPAI model becomes part of an AI system, and this integration imparts a general and multifunctional character to the system, it is then considered a GPAI system under the Act.

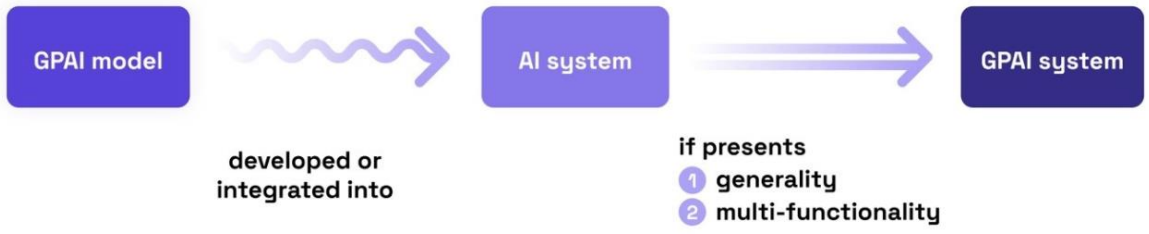

Figure 6. *GPAI Models and Systems, EU AI Act*

In line with the EU AI Act's risk-based approach, a GPAI may operate independently or be part of other AI systems, including those designated as high-risk. The Act's high-risk classification primarily considers the use-cases or functioning of AI systems, whereas the GPAI classification is about the model's generality and capabilities. Hence, GPAI is neither exempt from nor an alternative to the high-risk category. Consequently, the obligations of GPAI model providers are distinctly outlined and separate from those of high-risk AI system providers.

Under the legislation, GPAI models can be provided to downstream AI developers and users in different ways, namely through:

- API access, offering controlled model usage through provider-managed interfaces, where the control over the model and source code remains with the model provider (e.g. ChatGPT).
- Licensed open-source access, which enables access to GPAI models publicly under permissive licenses. Downstream developers can download, modify, and distribute the model under open-source access (e.g. Stability AI)

Obligations for the providers of general GPAI models range from drawing technical documentation, establishing copyright policies to providing detailed summaries for training data, among others. However, the EU AI Act provides for certain exemptions to such GPAI models made accessible to the public under free and open licenses. However, the scope of such licensing must satisfy the conditions below:

- Licensing must allow for public access, use, modification, and distribution of the model.
- The provider must make public the parameters of the model, including its weights, architectural details, and information on how the model is used.

Table 2 below maps these obligations and corresponding exemptions for models shared under free and open licenses.

Obligation	Description	Point of Reference for Content	Exemption for Open and Free License
Drawing up technical documentation	Providers must create and update technical documentation of the general-purpose AI model, including its training, testing process, and evaluation results.	Annex IXa of the EU AI Act (to be also facilitated via the Codes of Practice of the AI Office)	Granted
Providing information and documentation to AI system providers	Without jeopardizing their intellectual property rights and trade secrets, providers must create, update, and offer information and documentation to AI system providers intending to integrate the GPAI model. This documentation should help prospective AI system providers understand the model's capabilities and limitations in the observance of their obligations under the Act.	Annex IXb of the EU AI Act (to be also facilitated via the Codes of Practice of the AI Office)	Granted
Establish copyright policy	Providers must implement a policy to adhere to Union copyright law, especially in identifying and respecting rights reservations expressed under Article 4(3) of Directive (EU) 2019/790. They should employ state-of-the-art technologies to ensure compliance.	(Not specified)	N/A
Providing a detailed summary of training content and data	Providers must develop and publicly share a detailed summary about the content used for training the GPAI model. This summary should offer insight into the data and methodologies used during the training process.	AI Office template (to be developed by the AI Office)	N/A
Transparency for natural persons	If GPAI systems are designed to directly interact with natural persons, providers must structure their systems in a way that would inform the individuals that they are interacting with an AI system (unless it is reasonably obvious in light of the circumstances and context of use).	(Not specified)	Granted

Table 2. *Obligations and exemptions for models shared under free and open licenses in accordance with the EU AI Act.*

The legislation also exempts AI systems that are specifically developed and used for "the sole purpose" of scientific research and development. However, many observers have pointed out that this provision has created a loophole in which AI models originally produced for scientific research are then repurposed for commercial objectives evade the Act's other safety regulations.

Furthermore, the EU AI Act introduces a distinct category of risk within GPAI Models, known as General-Purpose AI Models with Systemic Risks (GPAISRs). These are defined as systems trained with a compute threshold exceeding 10^25 FLOPs, subject to potential revision by the European Commission. Obligations for GPAISRs are notably more rigorous, encompassing model evaluations, adversarial testing, cybersecurity measures, and mechanisms for promptly tracking, documenting, and reporting incidents to the proposed AI Office. Even if distributed under free and open licenses, GPAISR models are not exempt from these obligations.

Multilateral Efforts

Along with fellow G7 nations, the United States has also agreed to the International Guiding Principles on AI and a voluntary Code of Conduct for AI developers under the Hiroshima AI process. It establishes ethical guidelines for the development and deployment of advanced AI systems, such as foundation models. The Code encourages many of the principles found in the ideas of openness, including transparency, accountability, and respect for privacy and data protection principles throughout the AI lifecycle. It also promotes collaboration among stakeholders, including governments, industry, academia, and civil society, to address the ethical challenges posed by AI technology.

Self-regulatory Mechanisms

Beyond regulation and related frameworks, groups within the AI community have begun to develop self-regulatory mechanisms to address the risks of open foundation models. The Responsible AI Licensing (RAIL) Initiative has proposed Open and Responsible AI licenses (OpenRAIL) to address the limitations of traditional open-source licenses in adequately addressing the ethical and socio-economic implications of foundation models. OpenRAIL licenses intend to integrate openness with responsibility by allowing royalty-free access while embedding restrictions on usage in critical scenarios. However, to address potential misuse of the released AI model, OpenRAILs require downstream adoption of the use-based restrictions by later redistribution and derivatives of the AI model.

AI providers can also enhance their own models for better auditing through technical means. Models like the BLOOM Framework, which offer tools for data analysis and constant evaluation, exemplify proactive approaches to address issues such as data contamination and privacy risks. Furthermore, the implementation of interactive demos and transparent evaluations enables broader red-teaming efforts and enhances transparency in AI governance.

Recommendation: There is a clear need for a comprehensive approach that integrates voluntary standards, regulatory oversight, and international collaboration to effectively manage the risks and maximize the benefits of foundation models with widely available weights. By embracing initiatives such as the NIST AI RMF, fostering transparency, and ensuring

compliance with relevant regulations, we can promote the responsible deployment of AI technologies while safeguarding against potential harms. As such, we call for the NTIA to leverage its authority as the President's representative on related policy issues to facilitate various initiatives. For instance, the NTIA can drive alignment on the intersection of model release and AI safety through appropriate forums such as the U.S. Safety Institute, the annual AI Safety Summit and other multilateral convenings. Considering the NTIA's long running relationship with NIST, it is also well placed to collaborate with them on the development of interoperable standards and crosswalks. These would focus on model evaluations, safety testing, and external auditing linked to model release and open access. Lastly, the NTIA and NIST collaboration should seek partnerships with external experts to develop proportionate open model licensing regimes.

3. Concluding statement

Holistic AI welcomes the opportunity to provide comments on this important matter. We appreciate the open and collaborative approach taken by the NTIA, as it develops a report for the President on the benefits, risks, other implications and appropriate policy and regulatory approaches to dual-use foundation models with widely available model weights. We stand ready to support the NTIA and other public authorities or agencies involved in the development of responsible AI frameworks and resources.

Please contact publicpolicy@holisticai.com for any further information or follow-up on this submission.

Sincerely,

Holistic AI Inc.
publicpolicy@holisticai.com