Request for Comment (NTIA–2023–0009) on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

Organization: Future of Life Institute

Point of Contact: Hamza Tariq Chaudhry, US Policy Specialist. hamza@futureoflife.org

About the Organization

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Since its founding, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence, in 2017. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at Senate AI Insight Forums, participated in the UK AI Summit, and connected leading experts in the policy and technical domains to policymakers across the US government. We thank the National Telecommunications and Information Administration (NTIA) for the opportunity to respond to this request for comment (RfC) regarding (RfC) regarding Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights, as specified in the White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

EXECUTIVE SUMMARY

FLI has a long-standing tradition of thought leadership on AI governance, focusing on mitigating risks and maximizing the benefits of AI. In line with this mission, we have undertaken research and policy work to explore the potential risks, benefits, and policy approaches to governing various AI systems. Our work has included a particular focus on understanding the implications of "dual-use foundation models with widely available model weights" and developing appropriate governance strategies for these systems. This term was recently introduced in President Biden's Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI,[1] and will be referred to as Open Dual-Use Models (ODUMs) hereafter.[2] In our contribution, we offer analysis and recommendations on how to manage the unique use-cases, harms, and benefits of ODUMs.

1. ANALYSIS

ODUMs pose significant risks to society due to their unique characteristics, potential for misuse, and the difficulty of evaluating and controlling their capabilities. The following high-level summary outlines our risk analysis, which is expanded in the subsequent sections of our response.

1. Open weight AI systems, including, but not limited to ODUMs, are fundamentally distinct from traditional open source software (OSS).
2. ODUMs present *unique* risks in comparison to closed/proprietary systems, including the capacity to be fine-tuned by third parties to remove safeguards or augment dangerous capabilities.
3. AI systems claimed to be 'open' lie across a spectrum of access, each carrying different levels of benefits and risks.
4. The evaluation of capabilities of ODUMs is especially difficult, which is reason for caution against the release of more powerful systems.
5. "Open" AI systems are already nearly as powerful as closed frontier systems, and evidence suggests that they will become much more powerful over time.
6. ODUMs have already exhibited harms that, without appropriate intervention, will compound over time.

---

[1] https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[2] We also use the term "open AI systems," to refer to AI systems that have widely available or "open" components, but do not meet the thresholds for "dual-use foundation models with widely available model weights."

2. RECOMMENDATIONS

Based on the analysis above, we advocate for the following policy recommendations:

Updated Definitions and Thresholds

- The compute threshold for models to be considered presumptively dual-use should be reduced by an order of magnitude to 10^{25} integer or floating point operations (FLOPs), which would match the EU AI Act's threshold for general purpose AI models with high-impact capabilities. The threshold for additional safety evaluation should arguably be lower still if the model weights are made widely available or planned to be made widely available, especially if the model's training data includes CBRN, cyber offensive, or other sensitive expertise. These thresholds should be subject to periodic and as-needed reductions in the event technological advances reduce the compute necessary to produce similar capabilities.
- Relevant government agencies should adopt a less restrictive definition of 'dual-use foundation models' to ensure that current and future systems of concern remain in scope.

Auditing and Developer Responsibility

- ODUMs should undergo thorough testing and evaluation in secure environments appropriate to their level of risk prior to deployment.

- Developers should be legally responsible for ensuring that ODUMs are not accessed by actors barred from access by the government and should be legally responsible for performing all reasonable measures to prevent their models from being retrained to critically enable illegal activities.

- Whistleblower protections should be augmented to cover reporting on unsafe practices in development and/or planned deployment of unsafe ODUMs.

International Cooperation and Enforcement

- NTIA should encourage cooperation with other major AI powers, including strategic competitors.

Public Access and Power Concentration

- To address the tension between avoiding the concentration of power of AI and ensuring its safety and security, initiatives like the National Artificial Intelligence Research

Resource (NAIRR) should be pursued to create "public options" for AI development.

Risk Analysis

BACKGROUND

Over the last three decades, OSS has greatly benefited society, but the emergence of ODUMs has raised concerns about the risks associated with 'open-source' AI systems. While ODUMs share some similarities with OSS, they have the potential to be used for malicious purposes, such as generating disinformation, mass producing harmful synthetic content, automating cyberattacks, and even assisting in the creation of harmful biological weapons on a scale that far exceeds traditional OSS. As the capabilities of AI systems increase and the weights of powerful models become more widely available, these risks are likely to grow, requiring careful consideration and active management of their unique risks.

ANALYSIS

1. Open weight AI systems, including, but not limited to ODUMs, are fundamentally distinct from OSS. When debating the benefits, risks and policy prescriptions for AI systems, policymakers and experts have at times drawn an analogy with OSS, which masks important distinctions between the two. In the case of traditional OSS, access to the source code provides users and developers with a comprehensive understanding of the software's functionality and capabilities. However, this is not the case with open AI systems. Even if the source code of an AI system is made available, it typically does not provide complete information about the system's behavior and potential. The key determinants of an AI model's capabilities lie in its model weights, which are the result of extensive training on vast amounts of data. The complex interactions and relationships between the model weights within the neural network architecture make it difficult for humans to interpret and understand the model's decision-making process and potential outputs, even with access to the individual weight values. As a result, releasing model weights for AI systems confers less benefit when compared with releasing the source code of OSS, as it does not guarantee a complete understanding of the system's capabilities and potential consequences.

This problem is taken a step further in the case of ODUMs because they are generally more capable and are likely to have unforeseen capabilities. The potential for misuse of these models is significantly higher, as malicious actors could leverage their advanced capabilities for harmful purposes. Consequently, the release of ODUMs necessitates even greater caution and consideration of the associated risks. Additionally, while traditional OSS products can themselves be useful for general purposes, they are not inherently designed to perform a wide range of tasks. In contrast, many ODUMs are specifically designed to be highly generalizable

and capable of undertaking a broad spectrum of tasks without explicit programming. ODUMs can learn and improve their performance over time through exposure to data and interactions, making them significantly more powerful and potentially dangerous than traditional OSS. This generality, capability, and adaptability, combined with their ability to learn and evolve, distinguish ODUMs from traditional OSS and raise critical concerns about their potential misuse or unintended consequences.

2. ODUMs present unique risks in comparison to closed systems, including the capacity to be fine-tuned by third parties to remove safeguards or augment dangerous capabilities. While all frontier AI systems present some risks, ODUMs present unique potential for misuse. Presently, there are a small number of companies developing frontier dual-use systems. If there are security, privacy, or content issues with a "closed" system, the developer can patch the system, affecting all instances of it, or take it offline to address these issues. However, once model weights are available for download by the public, their release is irreversible, making reliably patching vulnerabilities top-down - or deploying guardrails around dangerous capabilities - impossible. Ultimately, like OSS, it is up to the downstream entities to choose to adopt updates. These entities may continue with the original, unpatched version, or may even fork the project to preserve or further enhance the dangerous capabilities.

Defining the openness of a model based on its level of distribution is a topic of ongoing debate. Some suggest restricting access to certain entities, such as academic institutions, while keeping the model open for others. However, enforcing such access restrictions is incredibly challenging in practice. Once model weights become widely available, there is currently no way to regain control, restrict access, or limit their use.

The release of model weights also allows for trivial removal of safeguards and lowers the barrier to entry for adapting systems toward more dangerous capabilities through fine-tuning.[3,4,5] Fine-tuning in the context of AI involves training an AI model on a specific task or domain to improve its domain- or use case-specific performance. AI systems whose model weights are made publicly available allow actors to directly access the model's weights without oversight and fine-

---

[3] Lermen, Simon, Charlie Rogers-Smith, and Jeffrey Ladish. "LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B." *arXiv*, Palisade Research, 2023, https://arxiv.org/abs/2310.20624.

[4] Gade, Pranav, et al. "BadLlama: Cheaply Removing Safety Fine-Tuning from Llama 2-Chat 13B." *arXiv*, Conjecture and Palisade Research, 2023, https://arxiv.org/abs/2311.00117.

[5] Yang, Xianjun, et al. "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models." *arXiv*, 2023, https://arxiv.org/abs/2310.02949.

tune it for their specific needs without having to train the model from scratch, which can be prohibitively expensive and therefore limited to a small number of well-resourced corporations. Fine-tuning enables the creation of specialized AI models tailored to specific domains or tasks while still benefiting from the general knowledge and baseline capabilities of the pre-trained model. Furthermore, the release of model weights, and the subsequent ability for third-party actors to build on top of them, expands the amount of actors to manage from a small handful of AI companies to potentially millions of independent actors, making it far more difficult to universally mitigate these harms.

Finally, in the absence of robust safeguards and comprehensive regulations governing the release of model weights, it is highly likely that we will see the emergence of AI systems that lack essential safety measures while possessing the full capabilities of advanced models. The primary reason why this has not yet happened with the most capable systems is due to the internal policies at top companies (e.g. OpenAI, Google Deepmind, Anthropic, etc.). If OpenAI made the decision to provide open access to their models, there would be no gap between the capabilities of open and closed AI systems. This scenario poses significant risks to society, as these unrestrained systems could be used for malicious purposes in the future or cause unintended harm on a large scale.

3. AI systems claimed to be 'open' lie across a spectrum of access, each carrying different levels of benefits and risks. In the last few years, a number of leading AI labs, including OpenAI, Anthropic, Meta, Google, Deepmind, among others, have released foundation models. While some models remain highly restricted, limiting who can access the model and its components, other models and their developers provide open access to model weights and architecture. For a detailed assessment on the risks and opportunities across the spectrum of access to ODUMs, see (Brammer 2023).[6]

Presently, many companies developing frontier systems are focused on the release of model weights while remaining 'closed' on other crucial aspects of their models. This allows companies to enjoy the marketing benefits of "open-source" without fully committing to the principles of transparency and collaboration that underpin the open-source movement. For a detailed exploration of this busines tactic, see Widder (2024), which explores the use of open weight

[6]Institute for Security and Technology. *How Does Access Impact Risk? Assessing AI Foundation Model Risk Along A Gradient of Access*. Dec. 2023, https://securityandtechnology.org/wp-content/uploads/2023/12/How-Does-Access-Impact-Risk-Assessing-AI-Foundation-Model-Risk-Along-A-Gradient-of-Access-Dec-2023.pdf.

release as a tactic for market consolidation. [7]

Furthermore, as Widder et al. suggest, the concentration of computational power among a few large technology companies raises significant concerns about the potential for ODUMs to exacerbate antitrust issues. Developing and deploying powerful AI models requires massive datasets and computational resources, which are increasingly controlled by a limited number of major players in the industry. The high costs associated with training and running inferences on large-scale AI models create a significant barrier to entry for smaller competitors. Additionally, the software systems used to optimize computational power for AI development are often designed for efficiency in proprietary hardware environments and are developed and governed by the same companies that sell access to computational resources and license AI models.

As a result, these large companies could potentially use their control over computational resources and optimized software to lock in the ecosystem of developers by making their AI models and APIs the de facto standard. This could effectively create a vendor lock-in scenario, making it difficult for developers to switch to alternative providers or develop their own competing solutions. The concentration of power in the hands of a few large technology companies, combined with the potential for vendor lock-in, raises significant antitrust concerns, such as reduced competition, stifled innovation, limited consumer choice, and the potential for abuse of market dominance through anti-competitive practices, and is fundamentally at odds with the values of traditional OSS.

4. The evaluation of capabilities of ODUMs is especially difficult, which is reason for caution against the release of more powerful systems. In the past year, we have witnessed that, despite extensive internal testing, even closed AI systems often contain significant bugs or security failures.[8,9] This issue is even more severe with ODUMs, as models with widely available weights can be fine-tuned after release to remove safeguards or augment dangerous

[7]Widder, David Gray and West, Sarah and Whittaker, Meredith, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI (August 17, 2023). Available at SSRN: https://ssrn.com/abstract=4543807 or http://dx.doi.org/10.2139/ssrn.4543807

[8] Burgess, M. (2023, November 29). *OpenAI's Custom Chatbots Are Leaking Their Secrets*. WIRED. https://www.wired.com/story/openai-custom-chatbots-gpts-prompt-injection-attacks/

[9] Porter, J. (2023, March 21). *ChatGPT bug temporarily exposes AI chat histories to other users*. The Verge. https://www.theverge.com/2023/3/21/23649806/chatgpt-chat-histories-bug-exposed-disabled-outage

capabilities, even if they underwent rigorous evaluations prior to release.

Given the wide range of novel approaches that malicious actors could use to co-opt ODUMs for malicious purposes, and the growing usefulness of such models for enabling, encouraging, or executing malevolent actions, evaluations should not merely test whether developers can identify risks in the unmodified model. Instead, they should assess whether the technical safeguards in place are robust enough to prevent malicious use above a reasonable, pre-defined, risk threshold, even if the model is modified to weaken or remove protective features or to augment latent capabilities. If a dual-use model cannot meet this standard, its weights should not be made publicly available.

While investing in tools to mitigate these risks is crucial, significant time will also be necessary to develop them properly. In the interim, it is essential to adopt a precautionary approach when deploying the AI systems that are plausibly disruptive, given the scale of potential harm and the irreversibility of model weight release.

5. "Open" AI systems are already nearly as powerful as closed frontier systems, and evidence suggests that they will become much more powerful over time. LLAMA-2, an open-source AI model developed by Meta, has already demonstrated superiority to GPT 3.5, a closed-source model created by OpenAI, on a number of capabilities assessments. In December, Mistral AI released the model weights to Mixtral 8x7B, claiming that is outperfomes LLAMA-2 and GPT-3.5 on "most standard benchmarks."[10] While ODUMs have thus far lagged behind closed systems, this is purely an artifact of top AI companies (e.g. OpenAI) electing not to open-source their most advanced models based on primarily profit-motivated internal policy decisions. There is no evidence to suggest that ODUMs will continue to lag behind closed systems in the long term. If the leading AI companies were to change their stance on open-sourcing their models, e.g., due to changes in market conditions or their assessment of self-interest, the landscape would be significantly different.

Recent developments in the AI market, such as Meta's announcement to build 'Open-Source' AGI, Elon Musk's lawsuit against OpenAI and the subsequent release of Grok's model weights, and Mistral's agreement with Microsoft to offer exclusive access to its most powerful AI system, underscore the volatile climate in which these decisions are made and the need for government

[10] Mistral AI Team. "Mixtral of Experts." *Mistral AI*, 11 Dec. 2023, https://mistral.ai/news/mixtral-of-experts/.

intervention.[11,12,13] There is no guarantee that corporations will maintain a consistent position on the level of access granted to their systems. Instead, they may choose to 'open' or 'close' their models at any point based on market advantage, with pronounced consequences for safety, national security, and American competitive advantage. Additionally, even if ODUMs currently lag behind the most powerful closed AI models, they are likely to become much more powerful over time as more actors build on top of them or elect to release model weights, amplifying potential risks in both scale and likelihood.

6. ODUMs have already exhibited harms that, without appropriate intervention, will compound over time

We can categorize the potential harms associated with ODUMs into three distinct classes:

Kinetic Harms	Attacks on Perception and Cognition	Magnification of High-Risk AI
Facilitate the development of components or conditions to physically harm targets.	Purposeful use of deception, influence, and manipulation, among others, to target individuals and institutions with the intent of capturing/disrupting levers of power or sow instability.	Generate powerful AI systems (e.g. recursive self improvement) that are highly-capable, misaligned, devoid of safeguards, or made to identify the vulnerabilities of other AI systems.

Open AI systems have already demonstrated the potential to facilitate attacks on perception and cognition and cause kinetic harm to society, particularly in the areas of cyberattacks,

[11] *Open Release of Grok-1*. https://x.ai/blog/grok-os

[12] In fact, Meta has explicitly expressed an intention to create Artificial General Intelligence (AGI) and 'open-source' it. While definitions of AGI vary, it is commonly assumed to imply systems that are better than even expert humans at virtually all cognitive tasks. If they succeed in that objective, this would immediately reverse the capabilities dynamics between open and closed systems, with open systems then outpacing their closed competitors. The imminence of this threat is bolstered by the unprecedented amount of compute Meta has amassed in pursuit of this goal. This emphasizes the need for governance measures beyond academic and industry norms and trends. See

[13] Heath, A. (2024, January 18). *Meta's new goal is to build artificial general intelligence*. The Verge. https://www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview

disinformation, and the proliferation of CSAM.[14,15] The UK National Cyber Security Centre found that AI systems are expected to significantly increase the volume and impact of cyber attacks by 2025, with varying degrees of influence on different types of cyber threats. While the near-term threat primarily involves the enhancement of existing tactics, techniques, and procedures, AI is already being used by both state and non-state actors to improve reconnaissance and social engineering. More advanced AI applications in cyber operations are likely to be limited to well-resourced actors with access to quality training data and expertise, but the increasing commoditization of AI-enabled capabilities will make improved tools (including ODUMs) available to a wider range of threat actors, including cybercriminals and state-sponsored groups.[16]

In the case of CBRN risks, as of early 2024, a significant portion of evidence suggests that open AI systems models function as instruments comparable to internet search engines in facilitating the procurement of information that could lead to harm.[17] However, as the reasoning, planning, and persuasion capabilities of these models continue to grow as models become more advanced, their potential to be misused by malicious actors is also expected to increase.

Leading AI researchers and companies are working on advanced, general-purpose AI systems that are expected to have deep, wide-ranging expertise and the ability to act in complex environments with relative ease. If these AI systems are released as open-source models without proper safeguards, they could quickly be modified by bad actors to remove any built-in safety measures. This could lead to harms spanning the categories mentioned above, and would cause a significant detriment to public security, as such AI systems could facilitate the planning and execution of harmful, illegal activities at unprecedented scales and complexities.

---

[14] CrowdStrike. *2024 Global Threat Report*. CrowdStrike, 2023, https://www.crowdstrike.com/global-threat-report/.

[15] Thiel, David, Melissa Stroebel, and Rebecca Portnoff. "Generative ML and CSAM: Implications and Mitigations." *Stanford Cyber Policy Center*, Stanford University, 24 June 2023, https://cyber.fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations.

[16] *The near-term impact of AI on the cyber threat*. (n.d.). https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

[17] Irving, Doug. "Red-Teaming the Risks of Using AI in Biological Attacks." *RAND*, 25 Mar. 2024, https://www.rand.org/pubs/articles/2024/red-teaming-the-risks-of-using-ai-in-biological-attacks.html.

Recommendations

To mitigate the risks mentioned above, we propose the following recommendations for consideration by NTIA:

Definitions and Thresholds

1. The compute threshold for models to be considered presumptively dual-use should be reduced by an order of magnitude to 10^{25} integer or floating point operations (FLOPs), which would match the EU AI Act's threshold for general purpose AI models with high-impact capabilities.[18]

The current definition of a "dual-use foundation model" in the Executive Order presumptively includes "any model that was trained using a quantity of computing power greater than 10^{26} integer or floating-point operations." We applaud the Executive Order's recognition that models beyond a certain capability threshold (using training compute as a proxy) present inherent risks due to their unpredictability, their capacity for emergent capabilities, and their wide range of potential uses, both malicious and otherwise. However, we recommend that this threshold be lowered to 10^{25} operations, aligning with the EU AI Act's threshold for general purpose AI models with high-impact capabilities. Today's cutting-edge models have already exhibited some capacity for dangerous misuse and malicious use, and none of these models have thus far reached the threshold of 10^{26} operations. This adjustment is particularly crucial, and should arguably warrant further reduction, for models whose weights are made widely available or are planned to be made widely available, as the potential for misuse, harm, and unmonitored adoption is higher in such cases, and post-hoc remediation of latent risks is not possible.

The definition should also allow for the possibility of lowering the compute threshold moving forward as algorithmic improvements enable more efficient training and increased capabilities require far less compute. By incorporating a mechanism for adjusting the threshold based on technological advancements, the definition can remain effective and relevant in the face of rapid progress. The following is a suggested modification to the Executive Order's definition:

"...any model that was trained using a quantity of computing power greater than 10^{25} *integer or floating-point operations, with the understanding that this threshold may be lowered as algorithmic improvements enable more efficient training and increased capabilities with less*

[18] "Article 52a: Classification of General-Purpose AI Models as General Purpose AI Models with Systemic Risk." *EU Artificial Intelligence Act*, Future of Life Institute, https://artificialintelligenceact.eu/article/52a/.

compute."

2. Relevant government agencies should adopt a less restrictive definition of 'dual-use foundation models' to ensure that current and future systems of concern remain in scope. At present, capabilities generally scale with the amount of compute used to train a model and the model's parameter size, making compute thresholds useful proxies for identifying dual-use AI systems in the absence of reliable capability benchmarks. However, the definition of "dual-use foundation models" should remain flexible and inclusive, suggesting the use of multiple sufficiency thresholds (rather than necessity thresholds) for qualifying as dual-use for a general-purpose AI system. These thresholds should include some that directly assess a variety of consequential capabilities and subject matters, with the ability to add new thresholds as benchmarks emerge and in the face of algorithmic improvements. To accommodate this more inclusive concept of dual-use and appropriately support the inclusion of new benchmarks and algorithmic improvements, a slight modification to the Executive Order's definition of "dual-use foundation models" (and subsequent definition of ODUMs) could be made as follows:

*"[...]an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; **and** is applicable across a wide range of contexts; ~~and~~ **or an AI model** that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, human rights, national public health or safety, or any combination of those matters, such…"*

Auditing and Developer Responsibility

3. ODUMs should undergo thorough testing and evaluation in secure environments appropriate to their level of risk. These secure environments could include secure enclaves, data cleanrooms, or Sensitive Compartmented Information Facilities (SCIFs). The government should conduct these assessments directly or delegate them to a group of government-approved independent auditors. When assessing these models, auditors must assume that a) built-in safety measures or restrictions could be removed or bypassed once the model is released, and b) the model could be fine-tuned or combined with other resources, such as datasets, educational materials, additional software, or other AI systems, potentially leading to the development of entirely new and unanticipated capabilities. Insufficient safeguards to protect against dangerous capabilities or dangerous unpredictable behavior should justify the authority to suspend the release of model weights, and potentially the system itself, until such shortcomings are resolved.

4. Developers should be legally responsible for ensuring that ODUMs are not accessed by actors barred from access by the government, and should be legally responsible for performing all reasonable measures to prevent their models from being retrained to

critically enable illegal activities. Previous executive orders have established requirements for record-keeping and investigations related to transactions involving foreign malicious cyber actors and other actors barred by the government. However, the nature of ODUMs make it nearly impossible for developers to fulfill these requirements. When model weights are made widely available, it becomes intractable for developers to maintain records of distribution, especially when foreign entities are involved. This inherent characteristic of ODUMs undermines the objectives and concerns outlined in previous executive orders (Executive Orders 13694, 13757, and 13984), as comprehensive record-keeping is not feasible.

Despite the challenges in meeting these record-keeping requirements, developers should still be held accountable if it can be proven that their models were used by actors barred by the government. In such cases, even if the developers did not directly engage in the malicious activities, they should be liable for the consequences of releasing their model weights without adequate safeguards or controls. Existing statutes on aiding and abetting are in many cases insufficient to cover these circumstances, as they generally require intent on the part of the developer to assist in the underlying crime. Since ODUMs provide incidental but foreseeable aide to foreign malign actors, new legal mechanisms must be created.

5. Whistleblower protections should be augmented to cover reporting on unsafe practices in development and/or planned deployment of unsafe ODUMs. These protections should be expanded to cover a wide range of potential whistleblowers, including employees, contractors, and external stakeholders who have knowledge of unsafe practices. This protection should include legal protection against retaliation, confidentiality, safe reporting channels, and the investigation of reports documenting unsafe practices. It is not presently clear whether existing whistleblower protections pertaining to consumer product safety would be applicable in these circumstances; as such, new regulations may be necessary to encourage reporting of potentially dangerous practices.

International Cooperation and Enforcement

6. NTIA should encourage cooperation with other major AI powers, including strategic competitors. The inherent portability of ODUMs implies that potential harms are highly unlikely to be confined to a developer's country of origin. International cooperation is critical for two primary reasons: first, to encourage effective enforcement of restrictions and regulations within our own jurisdiction. International cooperation ensures a level playing field and mitigates largely unfounded concerns that governance efforts could place domestic industries at a competitive disadvantage. Second, international cooperation would encourage the adoption of similar restrictions in other countries so that dangerous and uncontrollable ODUMs are not

released, regardless of their country of origin.

Public Access and Power Concentration

8. To address the tension between avoiding the concentration of power of AI and ensuring its safety and security, initiatives like the National Artificial Intelligence Research Resource (NAIRR) should be pursued to create "public options" for AI. The question of how to safely govern ODUMs, and AI systems more broadly, presents some tension between preventing the concentration of AI power in the hands of a few powerful private interests and mitigating the risk amplication that would arise from broader access to potentially dangerous systems.

One potential solution is for the U.S. to further invest in "public options" for AI. Initiatives like the National Artificial Intelligence Research Resource (NAIRR) could help develop and maintain publicly-funded AI models and infrastructure. This approach would ensure that access to advanced AI isn't solely controlled by corporate or proprietary interests, allowing researchers, entrepreneurs, and the general public to benefit from the technology while prioritizing safety, security, and oversight.

CLOSING REMARKS

We thank NTIA for the opportunity to provide comments on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights. We believe that if the risks outlined above are effectively addressed, the potential benefits of ODUMs can be realized while mitigating the associated dangers.