Dual-Use Foundation Models with Widely Available Model Weights

NTIA Report

JULY 2024

National Telecommunications and
Information Administration

United States Department of Commerce

Contents

Executive Summary

Overview

As stated by President Biden, “Artificial Intelligence (AI) holds extraordinary potential for both promise and peril.”[1] The development of increasingly advanced AI models, such as dual-use foundation models, has significantly heightened the potential risks and benefits of AI systems. Many developers provide limited or no public access to the inner workings of their advanced models, including their weights.[2] In contrast, some developers, such as Meta, Google, Microsoft, Stability AI, Mistral, the Allen Institute for AI, and EleutherAI,[3] have released models – though not always their most advanced models – with weights that are widely available (i.e., open to the public by allowing users to download these weights from the Internet or through other mechanisms).

Dual-use foundation models with widely available model weights (referred to in this Report as open foundation models) introduce a wide spectrum of benefits. They diversify and expand the array of actors, including less resourced actors, that participate in AI research and development. They decentralize AI market control from a few large AI developers. And they enable users to leverage models without sharing data with third parties, increasing confidentiality and data protection.

However, making the weights of certain foundation models widely available could also engender harms and risks to national security, equity, safety, privacy, or civil rights through affirmative misuse, failures of effective oversight, or lack of clear accountability mechanisms.

In October 2023 President Biden signed the Executive Order (EO) on “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” Noting the importance of maximizing the benefits of dual-use foundation models with widely available model weights while managing and mitigating the attendant risks, Section 4.6 of the EO tasked the Secretary of Commerce, acting through the Assistant Secretary of Commerce for Communications and Information and in consultation with the Secretary of State, with soliciting feedback “from the private sector, academia, civil society, and other stakeholders through a public consultation process on the potential risks, benefits, other implications, and appropriate policy and regulatory approaches related to dual-use foundation models for which the model weights are widely available.”

> **“Dual-use foundation models with widely available model weights...introduce a wide spectrum of benefits. They diversify and expand the array of actors, including less resourced actors, that participate in AI R&D. They decentralize AI market control from a few large AI developers. And they enable users to leverage models without sharing data with third parties, increasing confidentiality and data protection.”**

The EO further denoted that the Secretary of Commerce, through the Assistant Secretary of Commerce for Communications and Information and in consultation with the Secretary of State and heads of relevant agencies, would author a report to the President on the "potential benefits, risks, and implications of dual-use foundation models for which the model weights are widely available, as well as policy and regulatory recommendations pertaining to those models." In fulfilment of this tasking, the National Telecommunications and Information Administration (NTIA) published a public Request for Comment in February 2024 and received 332 comments in response.[4] NTIA further conducted extensive stakeholder outreach, including two public events gathering input from a range of policy and technology experts. This Report and its findings are based in large part on this feedback.

This Report provides a non-exhaustive review of the risks and benefits of open foundation models, broken down into the broad categories of Public Safety; Societal Risks and Wellbeing; Competition, Innovation, and Research; Geopolitical Considerations; and Uncertainty in Future Risks and Benefits. It is important to understand these risks as *marginal* risks—that is, risks that are unique to the deployment of dual-use foundation models with widely available model weights relative to risks from other existing technologies, including closed weight models and models that are not considered dual-use foundation models under the EO definition (such as foundation models with fewer than 10 billion parameters).

Finally, the Report considers under what circumstances the U.S. government should restrict the wide availability of model weights for dual-use foundation models. It evaluates a range of policy approaches, assessing their risks and benefits. And it concludes that, at the time of this Report, current evidence is not sufficient to definitively determine either that restrictions on such open-weight models are warranted, or that restrictions will never be appropriate in the future.

Instead, this Report suggests that the government should actively monitor a portfolio of risks that could arise from dual-use foundation models with widely available model weights and take steps to ensure that the government is prepared to act if heightened risks emerge. Specifically, we recommend that the government:

1. **Collect evidence through:**
    - a. Encouraging standards and – if appropriate – compelling audits, disclosures, and transparency for dual-use foundation models (including those without widely available model weights);
    - b. Supporting and conducting research into the safety, security, and trustworthiness of foundation models and high-risk models, as well as their downstream uses;
    - c. Supporting external research into the present and future capabilities and limitations of specific dual-use foundation models and risk mitigations; and

   d. Developing and maintaining a set of risk portfolios, indicators, and thresholds.

2. **Evaluate evidence through:**

   a. Assessing the lag time between developers introducing capabilities in leading proprietary models, and those same capabilities being made available in open models;

   b. Developing benchmarks and definitions for monitoring and potential action if deemed appropriate; and

   c. Maintaining and bolstering federal government expert capabilities in technical, legal, social science, and policy domains to support the evaluation of evidence.

3. **Act on evaluations through actions such as:**

   a. Restrictions on access to models; or

   b. Other risk mitigation measures.

4. **Keep open the possibility of additional government action.**

These recommendations support the ability of developers electing to make model weights widely available at this time, while bolstering the government's ability to monitor whether future models pose risks that indicate that it may be appropriate to limit model weight availability or apply other appropriate risk mitigation measures. The Report provides high-level guidance and considerations for this recommendation.

This Report provides relatively little insight to *future* releases of dual-use foundation models with widely available model weights; however, the recommended action to monitor risks would help the government determine the capabilities of future dual-use foundation models with widely available model weights. Without changes in research and monitoring capabilities, this dynamic may persist: any evidence of risks that would justify possible policy interventions to restrict the availability of model weights might arise only after AI models with those capabilities, closed or open, have been developed or released.

In summary, this Report outlines a cautious yet optimistic path that follows longstanding U.S. government policies supporting widespread access to digital technologies and their benefits, while nonetheless preparing for the potential future development of models for which an alternate approach may be justified.

Glossary

This Report uses the following definitions, many of which arise from the definitions in Executive Order 14110:

Artificial intelligence (AI) means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to:

- perceive real and virtual environments,
- abstract such perceptions into models through analysis in an automated manner, and
- use model inference to formulate options for information or action. 15 U.S.C 9401(3).[5]

An **AI model** is a component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.[6]

A **dual-use foundation model** means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and 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, national public health or safety, or any combination of those matters, such as by:

A. substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical biological, radiological, or nuclear (CBRN) weapons;

B. enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or

C. permitting the evasion of human control or oversight through means of deception or obfuscation[7]

Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities.[8]

A **dual-use foundation model with widely available model weights** (or, in this Report, an **open foundation model**) is a dual-use foundation model whose model weights have been released openly to the public, either by allowing users to download them from the Internet or through other mechanisms.

The term **foundation model** is used synonymously with the term dual-use foundation model in this Report. However, the term has been used more broadly in the AI community, notably without the "tens of billions of parameters" requirement.[9] Further, all foundation models do not necessarily display "dual-use" capabilities.

Model weights are "numerical parameters within an AI model that helps determine the model's output in response to inputs."[10] There are multiple types of weights, including the pre-trained model weights, weights from intermediate checkpoints, and weights of fine-tuned models.

Background

AI Model Weights

An AI model processes an input—such as a user prompt—into a corresponding output, and the contents of that output are determined by a series of numerical parameters that make up the model, known as the model's *weights*. The values of these weights, and therefore the behavior of the model, are determined by training the model with numerous examples.[11] The weights represent numerical values that the model has learned during training to achieve an objective specified by the developers. Parameters encode what a model has learned during the training phase, but they are not the only important component of an AI model. For example, foundation models are trained on great quantities of data; for large language models (LLMs) in particular, training data can be further decomposed into trillions of sub-units, called tokens. Other factors also play a significant role in model performance, such as the model's architecture, training procedures, the types of data (or *modalities*) processed by the model, and the complexity of the tasks the model is trained to perform.[12]

Some model developers have chosen to keep these weights guarded from the public, opting to control access through user-focused web interfaces or through APIs (application programming interfaces). Users or software systems can interact with these models by submitting inputs and receiving outputs, but cannot directly access the weights themselves. If a developer does decide to make a model's weights widely available, three important consequences arise.

First, once weights have been released, individuals and firms can customize them outside the developer's initial scope. For instance, users can fine-tune models on new data, such as text from a language or cultural context not included in the original training corpus.[13] Other techniques, such as quantization,[14] pruning,[15] and merging multiple models together, do not require new data. Customization techniques typically require significantly less technical knowledge, resources, and computing power than training a new model from scratch. The gap between the resources required to customize pre-trained models compared to training a full model will likely continue to widen.[16, 17] This accessibility afforded by open weights significantly lowers the barrier of entry to fine-tune models for both beneficial and harmful purposes. Adversarial actors can remove safeguards from open models via fine-tuning, then freely distribute the model, ultimately limiting the value of mitigation techniques.[18]

Users can also circumvent some of these safeguards in closed AI models, such as by consulting online information about how to 'jailbreak' a model to generate unintended answers (i.e., creative prompt engineering) or by fine-tuning AI models via APIs.[19] However, there are significantly fewer model-based safeguards for open-weight models overall.

Second, developers who publicly release model weights give up control over and visibility into its end users' actions. They cannot rescind access to the weights or perform moderation on model usage.[20] While the weights could be removed from distribution platforms, such as Hugging Face, once users have downloaded the weights they can share them through other means.[21] For example, the company Mistral AI publicly released Mixtral 8x7b, a dual-use foundation model with widely available model weights via BitTorrent, a decentralized peer-to-peer file sharing protocol which is designed specifically to evade control by one party.[22]

Finally, open model weights allow users to perform computational inference using their own computational resources, which may be on a local machine or bought from a cloud service. This localizability allows users to leverage models without sharing data with the developers of the model, which can be important for confidentiality and data protection (i.e., healthcare and finance industry). However, it also limits the capacity to monitor model use and misuse, in comparison to models that only allow API or web interface access.

Model size and use is an important factor when considering the effectiveness of legal means such as takedown requests in controlling the wide distribution of model weights. Large models and models that are used heavily are more likely to leverage commercial datacenter infrastructure than smaller or less frequently used models.

The Spectrum of Model Openness

This Report focuses on widely available model weights, but developers of dual-use foundation models can release their models with varying levels of openness.[23] Weights, code, training or fine-tuning data, and documentation can all be made available through multiple channels with varying types of restrictions.

Multiple layers of **structured access** can provide varying levels of access to different individuals at different times.[24] For example, access to model weights could be given to vetted researchers, but not to the general public. Model sharing can involve a **staged release**, where information and components are gradually released over time. This is done to allow time for safety research and for risks at one stage to become apparent before increasing access. The time scale for staged releases can vary, since “generally substantial sociotechnical research requires multiple weeks, months, and sometimes years.”[25] There are currently a wide range of AI **licenses** in use, which can be used by themselves or in conjunction with forms of structured access. Some licenses require the user or downloader to agree to use and redistribution restrictions, sometimes including behavioral or ethical guidelines, though they can be hard to enforce.[26]

Even developers of models that are not “open” can increase transparency and visibility through comprehensive documentation. **Model cards** are one method for describing a model’s technical details, intended uses, and performance on evaluation and red-teaming efforts.[27] Independent of whether the training data itself is widely available, information about the training dataset(s) can be distributed using **data sheets**, where developers can share the processes they used to train the model and any artifacts or procedures involved in human-in-the-loop training such as data annotation or reinforcement learning with human feedback instructions.[28]

These openness factors can and should be considered at all stages of the AI lifecycle, including post-deployment. For instance, a dual-use foundation model can be open at one stage of development and closed at another, such as a base model that is open but that is customized to create a downstream, closed consumer-facing system.

An Approach to Analysis of Marginal Risks and Benefits

As mandated by Executive Order 14110, this Report analyzes “the potential benefits, risks, and implications of dual-use foundation models for which the model weights are widely available.”[29] The assessment of policy options to address such models specifically, versus potential interventions to address risks more broadly, is the touchstone of our analysis. This Report will provide a broad assessment of the *marginal* risks and benefits of dual-use foundation models with widely available model weights. We define marginal risks and benefits as the additional risks and benefits that widely available model weights introduce compared to those that come from non-open foundation models or from other t echnologies more generally. Public commenters generally agreed that a marginal risk and benefit analysis framework is appropriate for our analysis.[30]

> **“The consideration of marginal risk is useful to avoid targeting dual-use foundation models with widely available weights with restrictions that are unduly stricter than alternative systems that pose a similar balance of benefits and risks.”**

The consideration of marginal risk is useful to avoid targeting dual-use foundation models with widely available weights with restrictions that are unduly stricter than alternative systems that pose a similar balance of benefits and risks. This does *not* mean that it is wise to distribute an unsafe open model as long as other equally unsafe systems already exist. Risks from open models and closed models should both be managed, though the particular mitigations required may vary. In some cases, managing the risk of open models may pose unique opportunities and challenges to reduce risk while maintaining as many of the benefits of openness as possible.

As the basis for generating policy recommendations for open foundation models, this Report assesses the marginal benefits and risks of harm that could plausibly be affected by policy and regulatory measures. Marginal benefits and risks, as assessed in this Report, meet the following conditions:

1. **There is a difference in magnitude between dual-use foundation models with widely available model weights as compared to such models without widely available weights.**
   - Risks and benefits arising equally from both dual-use foundation models with widely available model weights and closed-weight dual-use foundation models are not considered “marginal.”[31]
2. **The benefits or risks are greater for dual-use foundation models than for non-AI technologies and AI models not fitting the dual-use foundation model definition.**
   - Only risks and benefits that arise differently from dual-use foundation models and models that do not meet this definition (e.g., models with fewer than 10 billion parameters) are considered “marginal.”
   - Similarly, the risks and benefits that exist equally in both dual-use foundation models AI and other technological products or services (such as Internet search engines) are not considered “marginal.”

3. **The risks and benefits arise from models that will have widely available weights in the future over and above those with weights that have al-ready been widely released.**

   As discussed above, once model weights have been widely released, it is difficult to “un-release” them. Any policy that restricts the wide availability of dual-use foundation model weights will be most effective on models that have not yet been widely released.

   When deciding whether to restrict the availability of a specific future set of dual-use foundation models, it is important to consider whether those future models will present substantially greater marginal risks and/or benefits over existing models with widely available model weights.

   Not all policy options require restricting the wide availability of model weights. This consideration is most relevant for those policy options that require restricting the wide availability of model weights.

Risks and benefits that satisfy all three conditions are difficult to assess based on current evidence. Most current research on the capabilities of dual-use foundation models is conducted on models that have *already* been released. Evidence from this research provides a baseline against which to measure marginal risks and benefits, but cannot preemptively measure the risks and benefits introduced by the wide release of a future model. It can provide relatively little support for the marginal risks and benefits of *future* releases of dual-use foundation models with widely available model weights, except to the extent that such evidence supports a determination about the capabilities of future dual-use foundation models with widely available model weights. Without changes in research and monitoring capabilities, this dynamic may persist: Any evidence of risks that would justify possible policy interventions to restrict the availability of model weights might arise only after those AI models, closed or open, have been released.

> **“Without changes in research and monitoring capabilities, this dynamic may persist: Any evidence of risks that would justify possible policy interventions to restrict the availability of model weights might arise only after those AI models, closed or open, have been released.”**

Risks and Benefits of Dual-Use Foundation Models with Widely Available Model Weights

This section considers some of the marginal risks and benefits posed by open foundation models. This section overviews the main factors identified in the Executive Order, the comments submitted to NTIA for this Report, and existing literature. Neither the risks and benefits discussed here, nor the categories they are grouped into, should be considered comprehensive or definitive. Other reports identified in the Executive Order also overview some of these topics at greater length.

One limitation of this Report is that many AI models with widely available model weights—while highly capable—have fewer than 10 billion parameters, and are thus outside the scope of this Report as defined in Executive Order 14110. However, the number of parameters in a model (especially in models of different modalities, such as text-to-image or video generation models) may not correspond to their performance. For instance, advances in model architecture or training techniques can lead models which previously required more than 10 billion parameters to be matched in capabilities and performance by newer models with fewer than 10 billion parameters. Further, as science progresses, it is possible that this dynamic will accelerate, with the number of parameters required for advanced capabilities steadily decreasing.

These limitations, along with other factors, ultimately lead us to recommend that the federal government adopt a monitoring framework to inform ongoing assessments and possible policy action. Future assessments of the risks and benefits of open foundation models would benefit from an evidence base that includes a robust set of "leading indicators," or measures that can act as warning signs for potential or imminent risk that future open foundation models may introduce. Those leading indicators might include assessments of the capabilities of leading closed-weight foundation models (as similar behaviors and performance are likely to be found in open foundation models within months or years[32]) and other assessments of the evolving landscape of risks and benefits.

Open foundation model capabilities and limitations are evolving, and it is difficult to extrapolate their capabilities, as well as their impact on society, based on current evidence. Further, even if we could perfectly extrapolate model performance, quantifying the marginal risks and benefits is extremely difficult. For these reasons, our analysis favors taking steps to develop the evidence base and improve research techniques, as we address in our policy recommendations.

Public Safety

This section examines the marginal risks and benefits to public safety posed by dual-use foundation models with widely available model weights. As the AI landscape evolves, these risks, benefits, and overall impacts on public safety may shift. The policy recommendations section addresses these challenges.

"One limitation of this Report is that many AI models with widely available model weights—while highly capable—have fewer than 10 billion parameters, and are thus outside the scope of this Report as defined in Executive Order 14110. However, the number of parameters in a model (especially in models of different modalities, such as text-to-image or video generation models) may not correspond to their performance."

RISKS OF WIDELY AVAILABLE MODEL WEIGHTS FOR PUBLIC SAFETY

Dual-use foundation models with widely available model weights could plausibly exacerbate the risks AI models pose to public safety by allowing a wider range of actors, including irresponsible and malicious users, to leverage the existing capabilities of these models and augment them to create more dangerous systems.[33] For instance, even if the original model has built-in safeguards to prohibit certain prompts that may harm public safety, such as content filters,[34] blocklists,[35] and prompt shields,[36] direct model weight access can allow individuals to strip these safety features.[37] While people may be able to circumvent these mechanisms in closed models, direct access to model weights can allow these safety features to be circumvented more easily. Further, these actions are much easier and require fewer resources and technical knowledge than training a new model directly. Such actions may be difficult to monitor, oversee, and control, unless the individual uploads the modified model publicly.[38] As with all digital data in the Internet age, the release of model weights also cannot feasibly be reversed.

While users can also circumvent safeguards in closed AI models, such as by consulting online information about how to 'jailbreak' a model to generate unintended answers (i.e., creative prompt engineering) or, for more technical actors, fine-tuning AI models via APIs,[39] methods to mitigate these circumventions for an API-access system, such as moderating data sent to a model and incorporating safety-promoting data during fine-tuning, exist. These same mitigation strategies do not reliably work on AI models with widely available model weights.[40] Experimentation with available model weights, while often helpful for research to employ defenses against previously unknown attacks, can also illuminate new channels for malicious actors to exploit proprietary models because open models are easier to manipulate and can share properties with closed models.[41]

One mitigation that may work is using techniques including tuning a model on distinct objective functions and weakening its ability to produce dangerous information, prior to its weights being made widely available. However, we currently have limited technical understanding of the relative efficacy of different safeguards, and protections available to closed models might end up providing significant additional protection.[42]

This Report considers two discrete public safety risks discussed in relation to dual-use foundation models with widely available model weights: (a) lowering the barrier of entry for non-experts to leverage AI models to design and access information about chemical, biological, radiological, or nuclear (CBRN) weapons, as well as potentially synthesize, produce, acquire, or use them; and (b) enabling offensive cyber operations through automated vulnerability discovery and exploitation for a wide range of potential targets.

Chemical, Biological, Radiological, or Nuclear Threats to Public Safety

Widely available model weights could potentially exacerbate the risk that non-experts use dual-use foundation models to design, synthesize, produce, acquire, or use, chemical, biological, radiological, or nuclear (CBRN) weapons.

Open model weights could possibly increase this risk because they are: (i) more accessible to a wider range of actors, including actors who otherwise could not develop advanced AI models or use them in this way (either because closed models lack these capabilities, or they cannot “jailbreak” them to generate the desired information); and (ii) easy to distribute, which means that the original model and augmented, offshoot models, as well as instructions for how to exploit them, can be proliferated and used for harm without developer knowledge.

I. ACCESSIBILITY

This ease of access may enable various forms of CBRN risk. For instance, large language models (LLMs) can generate existing, dual-use information (defined as information that could support creation of a weapon but is not sensitive) or act as chemistry subject matter experts and lab assistants, and LLMs with open model weights specifically can be fine-tuned on domain-specific datasets, potentially exacerbating this risk.[43] However, the CBRN-related information open models can generate compared to what users can find from closed models and other easily accessible sources of information (e.g., search engines), as well as the ease of implementing mitigation measures for these respective threats, remains unclear.[44]

Open dual-use foundation models also potentially increase the level of access to biological design tools (BDT). BDTs can be defined “as the tools and methods that enable the design and understanding of biological processes (e.g., DNA sequences/synthesis or the design of novel organisms).”[45] Intentional or unintentional misuse of BDTs introduces the risk that they can create new information, as opposed to large language models’ dissemination of information that is widely available.[46] While BDTs exceeding the 10B parameter threshold are just now beginning to appear, sufficiently capable BDTs of any scale should be discussed alongside dual-use foundation models because of their potential risk for biological and chemical weapon creation.[47]

II. EASE OF DISTRIBUTION

Some experts have argued that the indiscriminate and untraceable distribution unique to open model weights creates the potential for enabling CBRN activity amongst bad actors, especially as foundation models increase their multi-modal capabilities and become better lab assistants.[48] No current models, proprietary or widely available, offer uplift on these tasks relative to open source information resources on the Internet.[49] But future models, especially those trained on confidential, proprietary, or heavily curated datasets relevant to CBRN, or those that significantly improve in multi-step reasoning, may pose risks of information synthesis and disclosure.[50]

III. FURTHER RESEARCH

Further research is needed to properly address the marginal risk added by the accessibility and ease of distribution of open foundation models. For instance, the risk delta between jailbreaking future closed models for CBRN content and augmenting open models, as well as how the size of the model, type of system, and technical expertise of the actor, may change these calculations remains unclear. Previous evaluations on CBRN risk may not cover all available open models or closed models whose weights could be made widely available.[51] Future analysis should distinguish between and treat separately each aspect of chemical, biological, radiological, or nuclear risks associated with open model weights.

Experts must also assess the amount open models increase this risk in the context of the entire design and development process of CBRN material. Information about how to design CBRN weapons may not be the highest barrier for developing them. Beyond computational design, pathogens, toxins, and chemical agents need to be physically generated, which requires expertise and lab equipment to create in the real world.[52] Other factors, such as the ease of attaining CBRN material, the incentives for engagement in these activities, and other mitigation measures—i.e., current legal prohibitions on nuclear, biological, and chemical weapons—also determine the extent to which open models introduce a substantive CBRN threat.

Offensive Cyber Operations Risks to Public Safety

Modifying an advanced dual-use foundation model with widely available model weights requires significantly fewer resources than training a new model and may be more plausible than circumventing safeguards on closed models. It is possible that fine tuning existing models on tasks relevant to cyber operations could further aid in conducting cyberattacks—especially for actors that conduct operations regularly enough to have rich training data and experimentation environments.[53]

FORMS OF ATTACKS

Cyber attacks that rely on dual-use foundation models with widely available model weights could take various forms, such as social engineering and spear-phishing, malware attack generation, and exploitation of other models' vulnerabilities.

First, open foundation models could enable social engineering (including through voice cloning and the automated generation of phishing emails).[54] Attacks could also take the form of automated cybersecurity vulnerability detection and exploitation.[55] The marginal cybersecurity risk posed by the wide distribution of dual-use foundation models may increase the scale of malicious action, overwhelming the capacity of law enforcement to effectively respond.

Cyber-attackers could also potentially leverage open models to automatically generate malware attacks and develop more sophisticated malware, such as viruses,[56] ransomware,[57] and Trojans.[58] For instance, one of Meta's open foundation models, Llama 2, may have helped cyber-attackers illicitly download other individuals' employee login credentials.[59]

Finally, actors can leverage open models to exploit vulnerabilities in other AI models, through data poisoning, prompt injections, and data extractions.[60]

FURTHER RESEARCH

The marginal cybersecurity risk posed by dual-use foundation models with widely available model weights remains unclear, and likely varies by attack vector and the preexisting capabilities of the cyber attackers in question.[61] For years, tools and exploits have become more readily accessible to lower-resourced adversaries, suggesting that foundation models may not drastically change the state of cybersecurity, but rather represent a continuation of existing trends. In the near term, the marginal uplift in capabilities for cyber attackers that widely available weights introduce to social engineering and phishing uses of foundation models may be the most significant of possible risks.[62] Closed foundation models and other machine learning models that can detect software vulnerabilities, alongside other cyber-attack tools, such as Metasploit, can also be found online for free, and play a critical role in adversary emulation.[63] Further, while open models could provide new instruments for performing offensive attacks, hackers may not want to invest time, energy, and resources into leveraging these models to update their existing techniques and tools.[64] The extent to which a particular dual-use foundation model with widely available model weights would meaningfully increase marginal risk is therefore uncertain.

When an AI system or tool is built using a foundation model with widely available model weights, the inclusion of the model could introduce unintentional cybersecurity vulnerabilities into the application as well.[65] Promisingly, it is more readily possible to prevent these types of harms – where the deployer of the model does not desire for the vulnerability to be present – than harms intentionally leveraged by the deployer of the model.[66] In line with the Cybersecurity and Infrastructure Security Agency's Secure by Design guidance, developers of AI models – whether open or closed source – can take steps to build in security from the start.[67]

BENEFITS OF WIDELY AVAILABLE MODEL WEIGHTS FOR PUBLIC SAFETY

The open release of foundation model weights also introduces benefits. Specifically, widely available model weights could: (a) bolster cyber deterrence and defense mechanisms; (b) propel safety research and help identify safety and security vulnerabilities on future and existing models; and (c) facilitate transparency and accountability through third-party auditing mechanisms.

Cyber Defense

Open foundation models can further cyber defense initiatives. For example, various cyber defense models, such as Security-BERT,[68] a privacy-preserving cyber-threat detection model, are fine-tuned versions of open foundation models.[69] These models and other systems built on open models provide security benefits by allowing firms, researchers, and users to use potentially sensitive data without sending this data to a third-party proprietary model for processing. Models with widely available weights also have more flexibility to be narrowly optimized for a particular deployment context, including through quantization, allowing opportunities for cost savings. Thus, several intrinsic technical benefits of openness allow a wider range of users to benefit from the value foundation models introduce to securing computing systems.[70] For instance, entities have published cyber-defense toolkits and created open-source channels for collaboration on cyber defense.[71]

Furthermore, any advances in dual-use foundation models' offensive cyber-attack capabilities may also strengthen defensive cybersecurity capabilities. If dual-use foundation models develop advanced offensive capabilities, those same capabilities can be used in securing systems and defending against cyberattacks. By detecting and addressing otherwise-undetected cybersecurity vulnerabilities, dual-use foundation models with widely available model weights could facilitate stronger cyber-defensive measures at scale.[72] Parity of licit access to models that have offensive cyber capabilities is also important for accurate adversary emulation, as advanced international cyber actors may incorporate such models into their own tradecraft. However, these benefits must be contextualized within the larger cyber defense landscape, as many developers perform their most effective cyber defense research internally.

Safety Research & Identification of Safety and Security Vulnerabilities

Widely available model weights can propel AI safety research. Open foundation models allow researchers without in-house proprietary AI models, such as academic institutions, non-profits, and individuals, to participate in AI safety research. A broad range of actors can experiment with open foundation model weights to advance research on many topics, such as vulnerability detection and mitigation, watermarking failures, and interpretability.[73]

Actors can also tailor safeguards to specific use-cases, thus improving downstream models. Creating external guardrails for dual-use foundation models can pose an abstract, under-specified task; actors that use open models for specific purposes can narrow and concretize this task and add on more targeted and effective safety training, testing, and guardrails. For instance, an actor that fine-tunes a foundation model to create an online therapy chatbot can add specific content filters for harmful mental health content, whereas a general-purpose developer may not consider all the possible ways an LLM could produce negative mental health information.

Open foundation models allow a broader range of actors to examine and scrutinize models to identify potential vulnerabilities and implement safety measures and patches, permitting more detailed interrogation and testing of foundation models across a range of conditions and variables.[74] This scrutiny from more individuals allows developers to understand models' limitations and ensure models' reliability and accuracy in scientific applications. An open model ecosystem also increases the availability of tools, such as open-source audit tooling projects, available for regulators to monitor and evaluate AI systems.[75]

Experimentation on model weights for research may also help propel alignment techniques. Llama 2, for example, has enabled research on reinforcement learning from human feedback (RLHF), though the underlying RLHF mechanism was first introduced by OpenAI, a closed model-weight company.[76] Open models will likely help the AI community grapple with future alignment issues. However, as models develop, this research benefit should be weighed against the possibility that open model weights could enable some developers to develop, use, or fine-tune systems without regard for safety best practices or regulations, resulting in a race to the bottom with negative impacts on public safety or national security. Malicious actors could weaponize or misuse models, increasing challenges to effective human control over highly capable AI systems.[77]

Auditing and Accountability

Weights, along with data and source code, are a critical piece of any accountability regime. Widely available model weights more readily allow neutral third-party entities to assess systems, perform audits, and validate internal developer safety checks. While access to model weights alone is insufficient for conducting more exhaustive testing, it is necessary for most useful testing of foundation models.[78]

Expanding the realm of auditors and allowing for external oversight regarding developers' internal safety checks increases accountability and transparency throughout the AI lifecycle, as well as public preparedness for harms. This is for three reasons.

First, the developer may be able to use information from external auditors about its model's robustness to improve the model's next iteration, and other AI developers may be able to benefit from this information to identify potential vulnerability points to avoid in future models.

"Weights, along with data and source code, are a critical piece of any accountability regime. Widely available model weights more readily allow neutral third-party entities to assess systems, perform audits, and validate internal developer safety checks."

Second, third-party evaluations can hold developers accountable for their internal safety and security checks, as well as downstream deployers responsible for which models they choose to use and how, which could improve accountability throughout the AI lifecycle. Of note, it may be difficult to implement such a third-party evaluation system due to differences in evaluations, the lack of ability to articulate how models can fail, and the scale of potential risks. Accessible model weights, alongside data and source code, facilitate oversight by regulatory bodies and independent researchers, allowing for more effective monitoring of AI technologies.[79]

Finally, a robust accountability environment may increase public trust and awareness of model capabilities, which could help society prepare for potential risks introduced by AI. The public can then respond to and develop resiliency measures to potential harms that have been demonstrated empirically. A foundation model ecosystem in which many models have widely available weights also further promotes transparency and visibility within the field, making it easier for the broader community to understand how models are developed and function.[80]

These community-led AI safety approaches could result in safer models, increased accountability, and improved public trust in AI and preparedness for potential risks. This transparency is vital for fostering trust between AI developers and the public and encourages accountability, as work is subject to scrutiny by the global community.[81]

Geopolitical Considerations

This section highlights the marginal risks and benefits related to the intersection of open foundation models and geopolitics. The availability of model weights could allow countries of concern to develop more robust advanced AI ecosystems, which, given the dual-use nature of foundation models, could pose risks to national security and public safety, and undercut the aims of U.S. chip controls. These countries could take U.S.-developed models and use them to enhance, perhaps substantially, their military and intelligence capabilities. There are also risks that may arise from the imposition of international standards that are not in line with U.S. values and commercial interests. On the benefits side, encouraging the availability of model weights could bolster cooperation with allies and deepen new relationships with developing partners.

GEOPOLITICAL RISKS OF WIDELY AVAILABLE MODEL WEIGHTS

Implications for Global Digital Ecosystem

The wide availability or restrictions of U.S. origin model weights could have unpredictable consequences for the global digital ecosystem, including by prompting some states to restrict open-source systems, causing further fragmentation of the Internet based on level of AI openness (i.e., a "splinter-net scenario").

More restrictive jurisdictions may also be better placed to set the terms of AI regulation and standards more broadly, even if this comes at the cost of innovation. States looking for how to regulate AI generally, including open models, may naturally look to imitate or adapt already existing regulatory approaches. Regulatory requirements would breed demand for standards. The United States would in this scenario pay a penalty in its ability to shape international standards on AI, even if it is still at a net advantage in setting standards.

Inconsistencies in approaches to model openness may also divide the Internet into digital silos, causing a "splinter-net" scenario. If one state decides to prohibit open model weights but others, such as the United States, do not, the restrictive nations must, in some way, prevent their citizens from accessing models published elsewhere. Since developers usually publish open model weights online, countries that choose to implement stricter measures will have to restrict certain websites, as some countries' websites would host open models and others would not.

Accelerate Dual-Use AI Innovation in Countries of Concern

Countries of concern can leverage U.S. model weights for their own AI research, development, and innovation, which, given the dual-use nature of foundation models, may introduce risk to U.S. national security and public safety.

GENERAL AI INNOVATION

Many U.S. open foundation models are more advanced than most closed models developed in other nations, including countries of concern.[82] Developers in these countries may choose to use U.S. open models instead of expending the resources or time necessary to build their own.

This development bolsters dual-use AI innovation in these countries and could help them create advanced technological ecosystems that they (1) may not have been able to build otherwise and (2) do not have to expend significant resources or time to create. Individual companies can allocate the funds they would have spent on up-front training to downstream product enhancements, improved dissemination tactics (e.g., marketing), and new products. Governments that would have subsidized AI training could also apply saved resources to other initiatives, such as talent cultivation, other technologies, and various other sectors that provide economic and security benefits.

Actors in these nations can also glean insights from open models about model architecture and training techniques, bolstering their long-term dual-use AI innovation.[83] Leading labs have already developed models based on Llama 2's architecture and training process with similar capabilities to Llama 2.[84]

This dual-use foundation model R&D could then accelerate global competition to build powerful AI-enabled national security applications and undercut the aim of U.S. export controls on semiconductors and related materials.

DUAL-USE IMPLICATIONS

Countries of concern could incorporate open foundation models into their military and intelligence planning, capabilities, and deployments. Due to the lower-cost nature of models with widely available model weights and the ability to operate these systems in network-constrained environments, it is easier for countries of concern to integrate dual-use foundation models in military and intelligence applications. Actors could experiment with foundation models to advance R&D for myriad military and intelligence applications,[85] including signal detection, target recognition, data processing, strategic decision making, combat simulation, transportation, signal jams, weapon coordination systems, and drone swarms. Open models could potentially further these research initiatives, allowing foreign actors to innovate on U.S. models and discover crucial technical knowledge for building dual-use models.

Some foreign actors are already experimenting with this type of AI military and intelligence research and applications. Since actors can inference on open model weights locally, U.S. developers cannot determine when and how countries of concern are using their model weights to research military and intelligence innovations. As open models improve over time, countries of concern could potentially use them to create even more dangerous systems, such as adaptive agent systems and coordinated weapon systems. Making model weights widely available may also allow countries of concern to coordinate their AI development, allowing them to develop specialized models depending on their needs and diversify spending. For example, a country of concern could invest in a specialized model to assist in creating CBRN weapons, or create models focused on misinformation and disinformation.

The U.S. government has already recognized the national security threat of countries of concern gaining access to powerful, dual-use leading models and worked to mitigate it through actions such as restrictions on advanced computing chips and semiconductor manufacturing equipment exports to certain nations. These controls attempt to limit adversaries access to computing power, which slows their development of advanced AI models related to military applications. Widely available model weights could allow strategic countries of concern to avoid U.S. policy instruments designed to hinder computational capabilities. Strategic countries of concern would not need to make the same investments in semiconductors and other hardware in order to produce similarly high performing AI models with the wide availability of model weights.[86] Thus, the open release of some foundation models may speed up the dual-use AI innovation and military application of AI by countries of concern, directly propelling the industry and national security capabilities that export controls and other U.S. actions aim to restrict.

The open release of some specific model weights could undermine U.S. technological leadership by distributing access to cutting-edge models—and the associated technological power—to foreign adversaries.

Some have argued that models accessible only through APIs pose fewer national security risks.[87] OpenAI commented that it had already taken steps to disrupt certain malicious nation-state affiliated actors who were using ChatGPT for cyber operations, which has been enabled in part through its use of a closed model distributed via an API.[88] Notably, the fact that these operations were conducted using closed models suggests that adversaries might used closed models even when open foundation models are available.

FURTHER RESEARCH

While open model weights could bolster AI military innovations in an untraceable, unregulated manner, the marginal risk open models pose to overall military and intelligence R&D, as well as the amount U.S. open models support country-of-concern AI development for military and civil technologies, is unknown. The trade-off between country-of-concern usage of open models compared to their own domestic national security research also remains uncertain and depends on the country.

GEOPOLITICAL BENEFITS OF WIDELY AVAILABLE MODEL WEIGHTS

Increase Global Use of U.S. Open Models

Widely available model weights could increase global adoption of U.S. origin models, thereby promoting the development of a global technology ecosystem around U.S. open models rather than competitors' open models. A burgeoning foundation model ecosystem does not appear overnight; creating the necessary infrastructure, such as datacenters and software ecosystems, talent cultivation systems, such as AI education and training, and even a creative, innovation start-up culture requires significant time, funding, and consideration. If the option to use open U.S. models exists, foreign states and non-state actors may not want to invest the time, money, and energy necessary to create their own foundation model market. This incentive structure could increase foreign adoption of U.S. origin open models, which are currently less powerful than proprietary models. Additionally, widespread use of U.S. open models would promote the United States' ability to set global norms for AI, bolstering our ability to foster global AI tools that promote the enjoyment of human rights.

Foster Positive Relationships across the Globe

Nurturing the open model ecosystem could foster positive relationships between the United States, allies and other countries that benefit from an open model ecosystem, such as developing partners, which may lead to enhanced international cooperation as well as the promotion of human rights through technological cooperation. Countries that have more open AI model ecosystems can more easily participate in international technological cooperation and crucial research exchanges, while, in turn, supporting their own open AI model ecosystems.

The safe, secure, and trustworthy deployment of AI requires coordination with allies and partners. Many U.S. allies have expressed an interest in maintaining an open AI ecosystem. Supporting an AI ecosystem that includes foundation models with widely available model weights could also bolster existing U.S. alliances, as well as potentially enhance partnerships that continue to mature. For example, economies that actively support their open model ecosystems include the Republic of Korea, Taiwan, Philippines, France, and Poland. These like-minded partners are critical in the current geopolitical landscape, and further cooperation with them in the realm of foundation models with widely available model weights would only serve to deepen ties and reinforce these mutually beneficial relationships.

By building relationships with other like-minded partners, U.S. allies can promote models with widely available model weights to third countries. This will become increasingly important as AI is adopted in global majority nations. The more the United States and its likeminded partners coordinate on creating a narrative that U.S. models with widely available model weights can spur innovation and promote competition, the more adoption of U.S. AI models will take place in developing countries. Coordination with like-minded partners will also be important to generally manage risks that arise from AI.[89] Moreover, if developing partners have access to U.S. open models, they may use open models from foreign countries of concern models less, if at all.

Promote Democratic Values in the Global AI Ecosystem

U.S. open model weights could steer the technological frontier towards AI that aligns with democratic values. The U.S. spearheading frontier AI development, and other nations building models on the foundation of U.S. open models, may increase the likelihood that the many cutting-edge AI technologies, along with correlated training techniques, safety and security protections, and deployment strategies, are built to uphold democratic values. However, we would note that it is not a given that AI applications built on open model weights will preserve democratic values.

Societal Risks and Well-Being

Dual-use foundation models with widely available model weights have the potential to create benefits across society, primarily through the access to AI capabilities that such models provide. At the same time, they also pose a substantial risk of causing harms to individuals and society. As noted above, our assessment of risk is tied to a framework of marginal risk: "the extent to which these models increase societal risk by intentional misuse beyond closed foundation models or pre-existing technologies."[90] Further, due to the relative novelty of dual-use foundation models, especially models that generate output in modalities beyond text (i.e., video and image generation), combined with known difficulties in accurate reporting for societal risks, precise estimates of the extent of these risks (especially the marginal risk of open foundation models over other models) are challenging to produce.[91]

The societal risks and benefits discussed in this section e xtrapolate from existing applications of open models that do not meet the parameter size criteria for this Report.[92] For example, there are no text-to-image models with widely available model weights with over 10 billion (10B) parameters available today, while there are multiple text generative based open models over the 10B parameter threshold.[93] Other developers of sophisticated closed-weight multi-modal models, such as SORA from OpenAI, have not publicly announced how many parameters they have.

> **"Dual-use foundation models with widely available model weights have the potential to create benefits across society, primarily through the access to AI capabilities that such models provide."**

It is also important to note that the content discussed in this section is not created in a vacuum. Both risks and benefits accrue due to how easily and widely accessible the tools for content creation are, as well as how the content is distributed. In the case of harmful content, some measure of that risk is dependent on how effectively platforms, content distributors, and others can prevent its widespread distribution. In the case of privacy and information security, there are still open questions as to how much the model "memorizes" from the data sets it was trained on and how much of that "memorization" contains personally identifiable information.[94] These risks are also embedded in our social systems. While a number of social risks and benefits arise from open foundation models, this section covers only a select few.

SOCIETAL RISKS OF WIDELY AVAILABLE MODEL WEIGHTS

CSAM & NCII

Models with widely available weights are already used today for AI-generated child sexual abuse material (CSAM), AI-generated non-consensual intimate imagery (NCII), and other forms of abusive content.[95] Such content may include the depiction of wholly fabricated individuals as well as specific individuals created using preexisting images of them. This content disproportionately affects women and teens, although any individual can be affected, and creates a hostile online environment that undermines equitable access to online services.[96] Such content, even if completely AI generated, can pose both immediate and long-term harm to its targets, especially if widely distributed, and creates a systemic risk across digital platforms for gender-based harassment and intimidation.[97, 98]

Open foundation models lower the barrier to create AI-generated CSAM and NCII. Creating such content using an open foundation model requires only a set of images to fine-tune the model, as opposed to creating a model from scratch. Open foundation models and downstream applications built from them, such as so-called 'nudifying' apps,[99] have made it easy to create with little to no cost individually targeted NCII and CSAM, which significantly enables both the production and distribution of AI-generated (but highly realistic) NCII and CSAM.[100] They also make it easier to distribute CSAM and NCII because there is no limit on the amount of content that can be created, making it possible to rapidly generate large amounts of AI-generated material. In contrast, closed model providers can more easily restrict or prevent the creation of CSAM and NCII through restrictions on prompts as well as on APIs.

Prior to the wide availability of AI systems, synthetic CSAM (e.g., non-photo based CSAM) primarily focused on non-realistic categories of material, such as anime-styled CSAM.[101] Open foundation models that include pornography or CSAM in their training data, as well as downstream implementations of open models that have been fine-tuned on CSAM or similar material, allow for the creation of AI-generated CSAM that is realistic to the point of being easily confused with non-AI-generated images and even based on real individuals.[102] NCII content specifically is often based on individuals who have never shared any form of nude images online. Creating such content prior to the release of generative AI models at the level of realism now achievable previously required both considerable skill with photo editing tools as well as a significant investment of time.

> **"Models with widely available weights are already used today for AI-generated child sexual abuse material (CSAM), AI-generated non-consensual intimate imagery (NCII), and other forms of abusive content."**

While the threat of NCII specifically is not unique to open foundation models (in one of the most publicized incidents to date—when AI-generated NCII images of singer Taylor Swift spread across the Internet in early 2024—the images were created with a closed generative model), since the emergence of open foundation models, researchers have documented significant increases in AI-created CSAM[103] and NCII. For example, the release of Stable Diffusion 1.5, an open model with 860 million parameters[104] (and which was revealed to have included documented CSAM in its training data),[105] enables the direct creation of CSAM, and fine-tuned versions of the model have been used in downstream apps.[106] This increase in harmful content enables producers to flood online platforms with enough content to overwhelm platform trust and safety teams and law enforcement's capacity to ingest and process CSAM reports.

Due to the sheer volume and speed of production they enable, the availability of open foundation models to create CSAM and NCII represents an increase in marginal risk over both existing closed foundation models and existing technologies.[107] The legal and regulatory system devoted to investigating and preventing the distribution of CSAM is not equipped to handle this influx of content.[108] As open foundation models become more advanced, this threat will likely increase.

The mass proliferation of CSAM and NCII also creates a substantial burden for women, teens and other vulnerable groups to live and participate in an increasingly online and digitized society. Further, proliferation of CSAM and NCII can discredit and undermine women leaders, journalists, and human rights defenders, and the implications of this harm extend beyond the individual to society and democracy at-large.[109] Again, these are risks that do not exist in a vacuum; the magnitude of harm in part depends on the ability to distribute content at scale.

Political Deepfakes

The capacity for malicious actors to use open foundation models to create convincing political deepfake content across a variety of modalities has introduced marginal risk to the integrity of democratic processes.[110] Actors are already generating and disseminating political deepfakes using downstream apps built upon both open and proprietary models. For example, deepfake audio recordings created with a proprietary voice cloning AI model emerged in January 2024, mimicking President Biden discouraging voters in the New Hampshire primary via robocalls.[111] This incident created immediate and widespread concerns about the potential impact on voter turnout, as well as what such incidents portended for upcoming elections and the democratic process as a whole.[112]

Internationally, campaigners supporting the re-election of Prime Minister Modi and other politicians in India are using open models to create synthetic videos and deliver personalized messages to voters by name, creating concerns that the public may not be able to discern the fake content from authentic material.[113] (At least some post-election accounts indicate that this concern failed to materialize and that generative AI enabled politicians to more easily communicate with voters in the 22 official languages of India.[114]) In 2022, a deepfake of Ukrainian President Volodymyr Zelensky circulated widely online, in which the false imitation of President Zelensky urges Ukrainian soldiers to lay down arms.[115]

In the absence of detection, disclosure, or labeling of synthetic political content, malicious actors can create deepfake videos or audio messages to unduly influence elections or enable disinformation campaigns.[116] Once released, they can be difficult to remove from the Internet, even after they have been verified as fake, in part due to the reluctance of social media platforms to remove this content. Deepfakes can also increase the "liar's dividend": skepticism and disbelief about legitimate content, contributing to pollution of the wider information environment.[117] This could undermine democratic processes by confusing voters and reducing the public's ability to determine fake events from actual ones.

As with concerns about CSAM and NCII, most of the open models capable of producing political deepfakes today have fewer than 10 billion parameters, but evidence does exist that political deepfake content has already been created and disseminated using open models under the 10B threshold. While deepfakes are a widespread source of concern, current dual-use foundation models with widely available model weights may not substantially exacerbate their creation or inflict major societal damage given the existing ability to create deepfakes using closed models, but, as open foundation models develop, this risk may increase.

Disinformation & Misinformation

Similar to the concerns described earlier regarding CSAM and NCII, the primary risks are tied to the low barriers to entry that open models may create for greater numbers of individuals, as well as to coordinated influence operations to create content and distribute it at scale.[118] Further, researchers have expressed concerns that the capabilities of generative LLMs may allow foreign actors to create targeted disinformation with greater cultural and linguistic sophistication.[119] As noted in the section on geopolitical considerations, the wide availability of open U.S. models could bolster dual-use AI innovation in countries of concern, which can enable them to develop more sophisticated disinformation campaigns.

DISINFORMATION

While the release of open foundation models raises concerns about the potential to enable disinformation campaigns by adversarial actors, assessments are mixed regarding whether open models, at least at current levels of capabilities, pose risks that are distinct from proprietary models. There is evidence that open foundation models, including LLMs, are already being used today to create disinformation-related content.[120] Disinformation researchers anticipate that generative models "will improve the content, reduce the cost, and increase the scale of campaigns; that they will introduce new forms of deception like tailored propaganda; and that they will widen the aperture for political actors who consider waging these campaigns."[121] As a consequence of an anticipated increase in the production of disinformation related content, one commenter expressed concerns that such content produced at significant enough of a scale would later be ingested by AI systems as training data, perpetuating its half-life.[122]

While many agree that open foundation models enable a larger range of adversarial actors to create disinformation, others dispute the importance of this assertion. Some researchers argue that the bottleneck for successful disinformation operations is not the cost of creating it.[123] Because the success of disinformation campaigns is dependent on effective distribution, key to evaluating marginal risk is whether the potential increased volume alone is an important factor, such that it may overwhelm the gatekeepers on platforms and other distribution venues. Some are skeptical that this is the case.[124] Carnegie researchers argue that not only has disinformation existed long before the advent of AI, but that generative AI tools may prove useful to researchers and others combating disinformation.[125]

MISINFORMATION

Unlike disinformation, which implies intentional malfeasance, misinformation encompasses factually incorrect information, or information presented in a misleading manner.[126] All foundation models are known to create and even help propagate factually incorrect content.[127, 128] Malicious actors may intentionally use models to create this information, and models can unintentionally produce inaccurate information, often referred to as "hallucinations."[129, 130] The marginal risks open foundation models pose in regards to misinformation are similar to those raised by CSAM, NCII, deepfakes and disinformation: they may lower the bar for individuals to create misinformation at scale and allow for more prolific distribution of misinformation. These impacts may exacerbate the disruption to the overall information ecosystem[131] and high volumes of misinformation may overwhelm information distributors' capacity to identify and respond to misinformation. However, some researchers argue that consumption of misinformation is limited to individuals more likely to seek it out and that foundation models do not substantially alter the amount or impact of this content online.[132]

One aspect of this marginal risk that necessitates further study is how individuals react to misinformation when it is directly outputted from an AI system (e.g., a chatbot) compared to consumed on social media or another platform. Little research to date exists that interrogates the consumption of misinformation directly from AI powered tools.[133] The majority of the public does not yet seem to understand generative AI's propensity for producing inaccurate information and may place undue trust in these systems; there have been well-publicized instances of lawyers, for example, relying on ChatGPT to assist with brief writing only to be surprised and embarrassed to find that the tool fictionalized legal sources.[134] In other instances, generative models have output untrue and potentially slanderous information about individuals, including public figures.[135]

Discriminatory Outcomes

Discrimination occurs when people are treated differently, solely or in part, based on protected characteristics such as gender, religion, sexual orientation, national origin, color, race, or disability.[136] Discrimination based on protected classes is, unfortunately, a widespread issue, impacting many groups by race, gender, ethnicity, disability, and other factors.[137] There has been substantial documentation of AI models, including open foundation models, generating biased or discriminatory outputs, despite developers' efforts to prevent them from doing so.[138]

Open foundation models may exacerbate this risk because, even if the original model has guardrails in place to help alleviate biased outcomes, downstream actors can fine-tune away these safeguards. These models could also be integrated into rights-impacting systems with little oversight and no means of monitoring their impact. Thus, it may be difficult to prevent open foundation models or their downstream applications from perpetuating biases and harmful institutional norms that may impact individuals' civil rights.[139] Bias encoded in foundation models becomes far more powerful when those models are used in decisional contexts, such as lending,[140] health care,[141] and criminal sentencing.

From a regulatory perspective, commercial actors that implement tools built on open foundation models will be subject to the same federal civil rights laws as those who leverage single-use models or proprietary foundation models.[142] The breadth of the potential impact and the current lack of clear determination regarding how to eliminate bias from all forms of AI models indicates that more research is needed to determine whether open foundation models substantially change this risk.

“...it may be difficult to prevent open foundation models or their downstream applications from perpetuating biases and harmful institutional norms that may impact individuals’ civil rights. Bias encoded in foundation models becomes far more powerful when those models are used in decisional contexts, such as lending, health care, and criminal sentencing.”

SOCIETAL BENEFITS OF WIDELY AVAILABLE MODEL WEIGHTS

Releasing foundation model weights widely also introduces benefits for society. Specifically, widely available model weights can: (a) support AI use for socially beneficial initiatives; (b) promote creativity by providing more accessible AI tools for entrepreneurial or artistic creation and expression; and (c) provide greater ability to test for bias and algorithmic discrimination.

Open Research for the Public Good

Making foundation model weights widely available allows a broader range of actor researchers and organizations to leverage advanced AI for projects aimed at improving public welfare. This approach democratizes access to cutting-edge technology, enabling efforts across healthcare,[143] environmental conservation,[144] biomedical innovations, and other critical areas that benefit society.[145, 146, 147] Scientists and researchers can tailor open models to better suit specific research needs or experimental parameters, enhancing the relevance and applicability of their work.[148] This customization capability is crucial for advancing scientific inquiries that require specialized models to analyze unique datasets or to simulate particular scenarios.[149]

AI Access for Entrepreneurs and Creatives

The wide availability of model weights can catalyze creativity and innovation, providing entrepreneurs, artists, and creators access to state-of-the-art AI tools.[150] By lowering barriers to access, a broader community can experiment with AI, leading to novel applications and creations. New entrants into the marketplace would not have to pay for a closed model, for example, to utilize the benefits of advanced AI systems. This democratization fuels a wide range of entrepreneurial ventures and artistic expressions, enriching the cultural landscape and reflecting a diverse array of perspectives and experiences.[151] This democratization can be particularly beneficial for small and medium-sized enterprises, which may otherwise face significant barriers to accessing more advanced AI systems.

Bias and Algorithmic Discrimination Mitigation

The ability to test for bias and algorithmic discrimination is significantly enhanced by widely available model weights. A wider community of researchers can work to identify biases in models and address these issues to create fairer AI systems. Including diverse communities and participants in this collaborative effort towards de-biasing AI and improving representation in generative AI is essential for promoting fairness and equity. mitigating bias in AI. Commenters have highlighted the importance of transparency and oversight enabled by open model weights in attempting to fight bias and algorithmic discrimination in foundation models.[152] As more companies, local governments, and non-profits use AI for rights-impacting activities, such as in healthcare, housing, employment, lending, and education, the need to better understand how these systems perpetuate bias and discrimination. There is a long history in the civil rights community of collaborative testing, and the wide availability of model weights enables that tradition to continue.[153]

Competition, Innovation, and Research

This section covers the marginal risks and benefits dual-use foundation models with widely available model weights may introduce to AI competition, innovation, and research.[154]

In traditional products, like cars or clothes, much of the price paid by the consumer goes toward producing that specific item. However, AI models, like television shows and books, are information goods. Training an advanced AI model requires a vast amount of resources, including financial resources,[155] but once trained, the model can be reproduced at a much lower cost.[156] "Vertical" markets of information goods like this reduce competition and lead to the dominance of a small number of companies. Successful companies can use their resources to produce higher quality products, driving out competitors and gaining even more market control.[157] Despite this rapid growth, markets related to AI foundation models risk potentially tending towards concentration that may lead to monopoly or oligopoly. The potential tendency toward monopoly or oligopoly partially derives from the structural advantages that already-dominant firms may be able to take advantage of as the technologies develop. A few companies spend vast amounts on the physical infrastructure to train foundation models, rendering it difficult for academics, smaller companies, nonprofits, and the public sector to keep pace.[158] The $2.6 billion requested in January 2023 over six years for the National Artificial Intelligence Research Resource (NAIRR) is significantly less than the over $7 billion that Meta expects to spend on GPUs alone this year; Facebook, a Meta company, states that it will have 350,000 H100 GPUs by the end of 2024, whereas leading universities have just hundreds.[159, 160, 161] Companies retaining proprietary control over the most advanced AI models, with the biggest companies making the largest investments by far, could continue to concentrate the economic power that derives from foundation models and hinder innovation more broadly.[162] The potential entrenchment of incumbent firms risks significant harm to competition.

> **"The ability to test for bias and algorithmic discrimination is significantly enhanced by widely available model weights. A wider community of researchers can work to identify biases in models and address these issues to create fairer AI systems."**

The effects that dual-use foundation models with widely available model weights may have on these dynamics is uncertain. Open model weights are unlikely to substantially impact the advanced foundation model industry, given constraints such as access to compute and other resources. However, even with just a few foundation models in the ecosystem, downstream applications may generally become more competitive.[163] With this caveat in mind, there are certain effects to competition, innovation, and research that can be associated with dual-use foundation models with widely available model weights.

RISKS OF WIDELY AVAILABLE MODEL WEIGHTS FOR COMPETITION, INNOVATION, AND RESEARCH

Perception of more market diversity than actually exists due to other factors

Widely available model weights are only one of many components in the gradient of openness of AI dual-use foundation models,[164] and their availability alone may be insufficient to bring about significant and long-lasting benefits.[165] The degree to which dual-use foundation models with widely available model weights may provide benefits to competition, innovation, and research is not fully clear, but the benefits seem more likely to be realized or maximized when additional conditions are in place to permit their full utilization. In particular, the benefits of these models may vary depending on whether other components of a model (e.g., training data, model architecture) and related resources (e.g., compute, talent/labor, funding for research) are also readily available.[166] Furthermore, vertical integration of the AI stack among a few key players could serve to bottleneck upstream markets, which may impact downstream use and applications of these models.[167]

Thus, there is a risk that the dual-use foundation models with widely available model weights—without additional components being made available or other work being undertaken—may create the perception of more competition. Without additional openness and transparency, it could seem as if there are more players throughout the AI supply chain while only a few companies still control most of the compute and human capital. For example, a small number of companies currently dominate the tech sector, having grown to prominence in an era that embraced open-source software for many purposes. With open source software, programmers make their work freely available to the community.[168] This open sharing of resources has added trillions to the global economy[169] and is a staple of software development today and widely supported worldwide.[170] But it has also been argued to increase global inequality.[171] Investing in open source models can be an optimal business model for companies in ways that might lead to further market concentration,[172] without necessarily reinvesting into the communities that contributed to the development of the technologies.[173] Businesses might create an open AI model to create a "first mover advantage," leading to wider adoption of their particular technology. In turn, this might push competitors out of the marketplace, support free public development of the companies' internal systems,[174] and create future licensing opportunities.[175] In the context of open source, for example, some commentators have noticed that "where some tech companies initially fought open source, seeing it as a threat to their own proprietary offerings, more recently these companies have tended to embrace it as a mechanism that can allow them to entrench dominance by setting standards of development while benefiting from the free labor of open source contributors."[176] A similar dynamic may occur in the AI context.[177] At the same time, "[t]he history of traditional open-source software provides a vision of the value that could result from the availability of open-weights AI models—including enabling greater innovation, driving competition, improving consumer choice, and reducing costs."[178]

BENEFITS OF WIDELY AVAILABLE MODEL WEIGHTS FOR COMPETITION, INNOVATION, AND RESEARCH

Lower Market Barriers to Entry

Dual-use foundation models with widely available model weights provide a building block for a variety of downstream uses and seem likely to foster greater participation by diverse actors along the AI supply chain.

While these models still require vast resources to train and develop, and the resources necessary to train leading models are likely to increase,[179] broadened access to model weights may decentralize the downstream AI application market. Open models can help: (i) businesses across a range of industries integrate AI into their services and (ii) lower the barrier to entry for non-incumbents to innovate downstream AI applications.

First, widely available model weights offer a significant advantage to businesses by enabling the development of innovative products and the customization of existing applications. These enterprises can also augment and fine-tune these models to fit seamlessly into their specific, sector-based products, enhancing the functionality and user experience of their applications.[180] A company could leverage these models to create a bespoke internal knowledge base, optimizing information retrieval and decision-making processes within the organization. Organizations can also control their own models, which gives them more control over sensitive data used in fine-tuning, the biases of systems, and more, as opposed to accessing models through an API, which may raise latency and privacy concerns. This control may be particularly pertinent to healthcare and education service providers. However, the level at which dual-use foundation models with widely available model weights could affect market concentration in upstream and specialized markets necessitates further examination.[181]

Second, open foundation models can help lower the barrier to entry for smaller actors to enter the market for downstream AI-powered products and services by reducing upfront costs that would have gone into model development or costs associated with paying a developer to use one,[182] and enabling competition against entrenched incumbents[183] (who may cut off API access to start-ups that pose a competitive threat), potentially reducing switching costs.[184] Start-ups can leverage these models in a variety of "wrapper" systems, such as chatbots, search engines, generative customer service tools, automated legal analysis, and more. Lowering the barrier to entry can allow smaller companies and startups to compete on a more even scale with better resourced competitors in downstream markets, thereby diversifying and decentralizing the concentration of power.[185] This benefit applies internationally as well—open foundation models contribute to international development and reducing the global digital divide, goals stated in the U.S.-led UN General Assembly resolution, "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development."[186]

Further, the diversification of the AI ecosystem through dual-use foundation models with widely available model weights may also allow communities to access AI systems when they would otherwise not be served by large AI companies (e.g., because serving smaller communities may be less economically viable or the interest too niche to legitimize financially).[187] This could strengthen the national AI workforce and foster the creation of specialized products and services that serve these communities in particular.[188]

Bolster AI Research and Development

Widely available model weights allow actors without access to the resources needed to train large models, such as non-profits and academics, to contribute more effectively to AI research and development.[189] This increased access both facilitates and diversifies AI research and development, and helps ensure that development of AI systems considers a diverse range of equities, perspectives, and societal impacts.[190, 191]

A broader range of actors with varying areas of expertise and perspectives can contribute to an existing model, collaborate, and experiment with different algorithmic solutions and increase independent research reproduction and validation.[192] Open foundation models operate as the foundational infrastructure to power a wide variety of products, which allows the developers of these models to benefit from community improvements, such as making inference more efficient.

These models could help facilitate research and development into safe, secure, and trustworthy AI (e.g., bias research using open models, greater auditing capabilities);[193] efficiency, scalability, and capability in AI (e.g., quantization and memorization);[194] and deployment of AI systems into different sectors or for novel use cases.[195] Models with open weights have spurred the development of AI model evaluation benchmarks.[196] These models could also allow for research that is generalizable to the foundation model space as a whole, and some research may require or otherwise benefit from deeper levels of access than closed models may offer.[197, 198] Shifting away from open research methods may also incur a cost to communities that currently operate on openness.[199] At the same time, there may be a risk that a reliance on open models could reduce incentives for capital intensive research.[200]

The net benefit to AI R&D from widely available model weights may be limited, however. R&D may also depend on other factors, such as access to computational resources[201] and to components such as training data.[202] Further, potential limitations related to the lack of user feedback and model usage fragmentation may impact the degree of innovation associated with dual-use foundation models with widely available model weights; in particular, as some academics note, “open foundation model developers generally do not have access to user feedback and interaction logs that closed model developers do for improving models over time” and “because open foundation models are generally more heavily customized, model usage becomes more fragmented and lessens the potential for strong economies of scale.”[203]

Disrupt AI Monoculture

These models could also help disrupt potential “algorithmic monoculture” by introducing alternatives to leading firms’ proprietary models for downstream deployers. While, as noted above, open model weights may not impact the very frontier of foundation model competition, they will likely increase the amount of models available to create downstream products. “Algorithmic monoculture” has been described as “the notion that choices and preferences will become homogenous in the face of algorithmic curation.”[204] In an algorithmic monoculture, the AI ecosystem comes to rely on one or a few foundation models for a vast range of downstream applications and diverse use cases; this homogeneity throughout the ecosystem could lead to technological risks, such as black boxing, methodological uniformity, and systemic failures,[205] and societal concerns, including persistent exclusion or downgrading of certain communities, centralized cultural power, and further marginalization of underrepresented perspectives.[206] Algorithmic monoculture can result from market concentration in a few select foundation models that impact a range of downstream applications and users.[207] Widely available model weights may mitigate algorithmic monoculture by allowing for greater algorithmic diversity. However, the extent of this mitigation to a monoculture effect may itself be affected by their own number and variety; for example, a dual-use foundation model with widely available model weights with sufficient adoption could itself create its own algorithmic monoculture based on the widely adopted model.

Uncertainty in Future Risks and Benefits

Many benefits and harms of foundation models are already occurring. However, some of these risks are yet speculative or unforeseen, while other risk/benefit areas can be identified. The potential future outcomes are so uncertain that effective, definitive, long-term AI strategy-setting is difficult. Indeed, some effects may be considered harms by some people and benefits by others.[208] While this is true for many of the benefits and risks discussed in this report, this observation particularly complicates classifying uncertain futures as beneficial or harmful, which can imply a level of confidence that does not exist.

Importantly, new effects arise not from foundation models alone, but from where and how technology interacts with people and systems, such as the economy or our social relationships.[209] Human creativity is a major driver of these new effects, and openness allows more human creativity to access dual-use foundation models. The economic impacts of AI, and in particular of large foundation models, become far more powerful when any company can tap into them, diversifying the number of use cases of these models.[210] As is the case with any new technology, societal risks can emerge. For example, the ability to use dual-use foundation models to create deepfakes is problematic when a foreign agent can use it to disrupt American elections,[211] but the risk is magnified when high school children everywhere can make a fake video of their classmates.[212] Bias encoded in foundation models becomes far more powerful when those models are used to determine lending,[213] health care,[214] and prison terms.[215] AI has more effects when it is more capable *and* when it interacts with more people and systems. Technological testing and evaluation are useful for examining models based on technical capabilities, but cannot anticipate the many ways that a model might be used when set loose in society.

Openness tends to increase the number of new effects and uses of technologies, including foundation models. Thus, open foundation models will, generally speaking, likely increase both benefits and risks posed by foundation models. However, it is important to note that there is significant uncertainty around the harms/benefits of any specific use,[216] and closed models carry their own unique benefits and risks.

Though dual-use foundation models are relatively new technologies, the challenge of adapting to unknown technological effects is not. As the World Wide Web became popular, proponents claimed it was a place to collaboratively come together.[217] Those proponents did not anticipate that this connection could, ironically, also lead to loneliness.[218] Advanced AI technologies are a particularly challenging policy problem, because AI develops so quickly[219] and business goals, rather than public interest, largely drive innovation.[220] Future policymakers will need to monitor risks and be adaptive as technology and society changes.

Methods for managing uncertain problems like advanced AI have been studied under a variety of frameworks.[221] Approaches to deal with the deep uncertainty around AI specifically include broad stakeholder participation,[222] explicit and repeated evaluation of values used for decision-making,[223] a focus on identifying and understanding hidden or potentially emergent issues to inform policymakers,[224] mapping out potential futures and scenarios,[225] and setting up dynamic plans involving potential actions, evaluations, and timelines for reevaluation under changing conditions.[226]

Summary of Risks and Benefits

Wide availability of open model weights for dual use foundation models could pose a range of marginal risks and benefits. But models are evolving too rapidly, and extrapolation based on current capabilities and limitations is too difficult, to conclude whether open foundation models, overall, pose more marginal risks than benefits (or vice versa), as well as the isolated trade-offs in specific sections. For instance, how much do open model weights lower the barrier to entry for the synthesis, dissemination, and use of CBRN material? Do open model weights propel safety research more than they introduce new misuse or control risks? Do they bolster offensive cyber attacks more than propel cyber defense research? Do they enable more discrimination in downstream systems than they promote bias research? And how do we weigh these considerations against the introduction and dissemination of CSAM/NCII content? The following policy approaches and recommendations consider these uncertain factors and outline how the U.S. government can work to assess this evolving landscape.

> **“Wide availability of open model weights for dual use foundation models could pose a range of marginal risks and benefits. But models are evolving too rapidly, and extrapolation based on current capabilities and limitations is too difficult, to conclude whether open foundation models, overall, pose more marginal risks than benefits.”**

Policy Approaches and Recommendations

The U.S. government could pursue a range of approaches to governing the risks and benefits of dual-use foundation models with widely available model weights. This Report considers three main policy approaches:

1. **Restrict the availability of model weights for dual-use foundation models**
2. **Continuously evaluate the dual-use foundation model ecosystem and build & maintain the capacity to effectively respond**
3. **Accept or promote openness**

This Report analyzes the pros and cons of these three approaches and ultimately concludes, as NTIA's recommendation, that the government should **not** restrict the wide availability of model weights for dual-use foundation models at this time. Instead, the U.S. government should actively monitor and maintain the capacity to quickly respond to specific risks across the foundation model ecosystem, by collecting evidence, evaluating that evidence, and then acting on those evaluations. The government should also continue to encourage innovation and leading international coordination on open models, while preserving the option to restrict the wide availability of certain classes of model weights in the future.

Policy Approaches

1.

RESTRICT THE AVAILABILITY OF MODEL WEIGHTS FOR DUAL-USE FOUNDATION MODELS

The U.S. government could seek to restrict the wide availability of model weights for specific classes of dual-use foundation models through existing authorities or by working to establish new authorities. Restrictions could take a variety of forms, including prohibitions on the wide distribution of model weights, controls on the exports of widely available model weights, licensing requirements for firms granted access to weights, or the limiting of access to APIs or web interfaces. A *structured access* regime would determine who can perform specific tasks, such as inference, fine-tuning, and use in third-party applications.[227] Another approach could involve mandating a *staged release*, where progressively wider access is granted over time to certain individuals or the public as the developer evaluates post-deployment risks and downstream effects.[228] Additionally, a government agency could require review and approval of model licenses prior to the release of model weights or at other stages in a structured access or staged release regime.

Pros: Proponents of restricting model weights argue that such measures are essential for limiting nefarious actors' ability to augment foundation models for harmful purposes. For instance, restrictions could reduce the accessibility of specific models trained on biological data, possibly creating a higher barrier to entry for the design, synthesis, acquisition, and use of biological weapons.[229] Additionally, limiting the availability of specific advanced open-weight models could potentially limit the ability of countries of concern to build on these models and gain strategic AI research advantages.[230] Restricting the wide availability of model weights could potentially limit the capabilities of countries of concern, as well as non-state actors, from developing and deploying sophisticated AI systems in ways that threaten national security and public safety.

Cons: Restrictions on the open publication of model weights would impede transparency into advanced AI models.[231] The degree of this effect, and other negative effects in this section, depend on the types and magnitude of restrictions. Model weight restrictions could hinder collaborative efforts to understand and improve AI systems and slow progress in critical areas of research, including AI safety, security, and trustworthiness, such as bias mitigation and interpretability.[232] Restrictions might also hamper research into foundation models, and stifle the growth of the field.[233] This could force investment and talent to relocate to more permissive jurisdictions, enhance adversary and competitor capabilities, and limit U.S. and allied autonomy to control the distribution of specific model weights. Targeted restrictions on certain classes of models may impose less of these costs than broader restrictions. Restrictions that use specific benchmarks or are not carefully scoped may not address some key risks and concerns.

> **"Restrictions on the open publication of model weights would impede transparency into advanced AI models...Model weight restrictions could hinder collaborative efforts to understand and improve AI systems and slow progress in critical areas of research, including AI safety, security, and trustworthiness, such as bias mitigation and interpretability."**

For instance, AI-generated CSAM and NCII are created using models with widely available model weights that are well below the 10 billion parameter threshold of a dual-use foundation model.[234] Further, if other countries with the current or future capacity to develop dual-use foundation models do not similarly restrict the wide availability of model weights, the risks will persist regardless of U.S. policy. Some commenters have argued that the sharing or open dissemination of model weights would be protected under the First Amendment, similar to protections that have been recognized by some courts for open-source software.[235]

2.

CONTINUOUSLY EVALUATE THE DUAL-USE FOUNDATION MODEL ECOSYSTEM AND BUILD & MAINTAIN THE CAPACITY TO EFFECTIVELY RESPOND

A second approach would require the U.S. government to build the capacity to continuously evaluate dual-use foundation models for evidence of unacceptable risk, and to bolster its capacity to respond to models that present such risk. The U.S. government can leverage the information and research that an open environment fosters to engage in ongoing monitoring of potential risks of dual-use foundation models. By staying up-to-date on model advancements, the U.S. government can respond to current and future risks in an agile and effective manner.

Effective risk monitoring would require access to information on both open and proprietary foundation models, including dual-use foundation models and other advanced AI models, systems, and agents.[236] Useful risk evaluation information could include data from foundation model developers, AI platforms, independent auditors, and other actors in the foundation model marketplace,[237, 238] model evaluations and red-teaming results,[239, 240] and standardized testing, evaluations, and risk benchmarks.[241, 242] It could also include keeping track of key indicators in the economic and social systems that impact and interact with foundation models.

Best practices around evaluation and transparency will change over time, as will society's perceptions of the most pressing risks, so the U.S. government would need flexibility in future adaptations of evaluation standards and transparency requirements. Monitoring of specific risks, such as CBRN or cybersecurity risks, may require liaising between agencies with specific subject matter expertise.[243] In addition, monitoring requires secure storage of the research, including for external research and internal research with proprietary data.[244] The risks that arise from open and closed foundation models involve not just the technology itself, but how those models interact with social, legal, and economic systems post-deployment.[245, 246] Consequently, effective monitoring and responsiveness would require combined technical, social, legal, and economic expertise.

Research and evaluation methods would need to be developed, including benchmarking, evaluation of capabilities, risks, limitations, and mitigations, red-teaming standards, and methods for monitoring and responding when appropriate to the more social, long-term, and emergent risks. International cooperation would also be needed.[247] As other nations develop their governance frameworks for foundation models, the U.S. could work to collaborate on interoperable standards and guidelines with like-minded partners.

Pros: A monitoring approach gives time for the U.S. government to develop the staffing, knowledge, and infrastructure to respond to AI's rapid developments.[248] Monitoring allows for a more targeted approach to risk mitigation. If done well, it allows the United States to continue to benefit from the wide availability of model weights, such as through innovation and research, while protecting against both near- and long-term risks. The uses of AI will likely continue to change, as will the technology itself, and the marketplace of model developers, distribution platforms, companies using fine-tuned models, and end users.[249] A monitoring approach would give time for the U.S. government to develop the staffing, k nowledge, a nd i nfrastructure to respond appropriately.[250] In addition, the increased AI capabilities that could come from this approach could support continued U.S. leadership on the international AI front.

> **"Monitoring allows for a more targeted approach to risk mitigation...[Monitoring] allows the United States to continue to benefit from the wide availability of model weights, such as through innovation and research, while protecting against both near- and long-term risks."**

Cons: Besides the potential risks of not restricting open model weights mentioned above, such as enabling innovation in countries of concern, one major drawback is the cost to the U.S. government. AI will impact many corners of government, so cross-sector monitoring capacity will likely require significant investment. Monitoring imposes obligations on companies, which could be costly, especially for smaller companies in the AI value chain, and burden the U.S. innovation ecosystem. Compelled disclosures to the government and public could also be intrusive and would need to be carefully considered to avoid exposure of proprietary information. If this approach is not done well, it could be a drain on government expenditures without substantially mitigating risks. For example, as innovation leads to new uses, more unexpected harms will likely arise that require a government response. The U.S. government may also incur extra financial mitigation costs in areas such as cybersecurity defense.

3. ACCEPT OR PROMOTE OPENNESS

The U.S. government has tended toward a laissez-faire approach to many new technologies in order to promote innovation and permit market forces to shape the development of technology.[251] On the one hand, a hands-off approach to the wide availability of dual-use foundation model weights can enable different competitive approaches to the development of foundation models[252] but would rely on industry and the research community to develop methods for detecting and mitigating risks. Several foundation model developers have already articulated risk detection and mitigation frameworks that could serve as the focus for broader norm development across the industry.[253] On the other hand, the U.S. government could further affirmatively promote the wide availability of model weights for dual-use foundation models. Further active steps could be taken, for example government policy could be used to support open foundation models through subsidies, procurement rules, or regulatory support for open models.

Pros: An approach involving minimal government action would pose the least risk of regulatory burden on developers of dual use foundation models. It is likely that the main benefits of openness would arise from innovation and research.[254] Openness may provide more access for small businesses to access foundation model resources.[255] Open resources are the norm among academic researchers, who draw on previous work to build a collective, public body of knowledge.[256] In recent years, private companies have overtaken academics in AI research.[257, 258] Encouraging openness could potentially reverse that trend. In addition, incentives for openness could support greater access for researchers to examine models for safety, security, and trustworthiness, including bias and interpretability.[259]

Cons: There are several significant drawbacks to a hands-off or affirmative promotion approach. There has already been significant involvement by both the U.S. and other allied governments in obtaining industry commitments and developing standards for AI risk management. Also, as discussed, there are significant security, societal, and strategic risks that may yet materialize from dual-use foundation models. This option would constrain the ability of the U.S. government to understand the developing risk landscape or to develop mitigation measures. Incentivizing openness may well exacerbate many of the risks from dual-use foundation models with widely available model weights that have been outlined in this Report.[260] For example, without restrictions on sharing model weights, dual-use foundation models that create novel biorisk or cybersecurity threats could be used by a wide range of actors, from foreign nations to amateur technologists. As innovation leads to new uses, new and unexpected harms will likely arise. Besides the negative societal effects that these risks could create, the U.S. government may also incur extra financial mitigation costs in areas such as cybersecurity defense.

Recommendations

NTIA recommends that the federal government actively monitor and maintain the capacity to quickly respond to specific risks across the foundation model ecosystem, by collecting evidence, evaluating that evidence, and acting on those evaluations. The government should also continue encouraging innovation and leading international coordination on topics related to open foundation models. This recommendation preserves the option to restrict the wide availability of certain future classes of model weights if the U.S. government assesses that the risks of their wide availability sufficiently outweigh the benefits (option 1). This will allow the federal government to build capacity to engage in effective oversight of the ecosystem and to develop a stronger evidence base to evaluate any potential interventions in the future.

As of the time of publication of this Report, there is not sufficient evidence on the marginal risks of dual-use foundation models with widely available model weights to conclude that restrictions on model weights are currently appropriate, nor that restrictions will never be appropriate in the future. Prohibiting the release of some or all dual-use foundation model weights now would limit the crucial evidence-gathering necessary while also limiting the ability of researchers, regulators, civil society, and industry to learn more about the technology, as the balance of risks and benefits may change over time.

Active monitoring by the federal government of the continued risks arising from dual-use foundation models with widely available model weights should involve a risk-specific risk management approach that includes three steps:

1. **Collect evidence** about the capabilities, risks, and benefits of the present and future ecosystem of dual-use foundation models with widely available model weights, and monitoring specific model-based and downstream indicators of risk for potential risk cases, as well as the difference in capabilities between open foundation models and proprietary models;
2. **Evaluate that evidence** by comparing indicators against specified thresholds, to determine when risks are significant enough to change the federal government's approach to open-weight foundation model governance; and when appropriate,
3. **Act on those evaluations** by adopting policy and regulatory measures targeted appropriately across the AI value chain.

The United States government does not currently have the capacity to monitor and effectively respond to many of the risks arising from foundation models. A significant component of our recommendation is to increase the government's capacity for evidence gathering, agile decision-making, and effective action.

STEP 1

Collect Evidence

We recommend that the federal government take steps to ensure that policymakers have access to a high-quality evidence base upon which to assess policy approaches to dual-use foundation models with widely available model weights going forward.[261] To develop and promote that evidence base, the federal government should:

A. Encourage, Standardize, and, if Appropriate, Compel Auditing and Transparency for Foundation Models

It is difficult to understand the risks of dual-use foundation models without transparency into AI model development and deployment, including downstream uses. To the extent reasonable, the federal government should standardize testing and auditing methods, which may vary based on the capabilities, limitations, and contexts of use of particular models and systems. The capabilities and limitations of closed-weight foundation models are currently good indicators of the potential future capabilities and limitations of open-weight models. The federal government should encourage, and where appropriate and where authority exists, require either independent or government audits and assessments of certain closed-weight foundation models[262] – especially closed-weight models whose capabilities exceed those of advanced dual-use foundation models with widely available model weights and can therefore serve as a leading indicator of the future capabilities of those models. For instance, the U.S. AI Safety Institute, housed in the National Institute for Standards and Technology, plans to perform pre-and post-deployment safety tests of leading models. This work should help the federal government understand and predict the risks, benefits, capabilities, and limitations of dual-use foundation models with widely available model weights. The federal government should also aim to enable independent researcher access, in addition to U.S. AI Safety Institute access, to certain closed-weight foundation models, including downstream effects of AI on the information individuals receive and how it affects their behavior. This will help assess the risks and benefits that could arise from future models.

As model capabilities and limitations change, so will the appropriate testing and auditing procedures. The United States should stay actively engaged in updating those methods and procedures. The federal government should establish criteria to define the set of dual-use foundation models that should undergo pre-release testing before weights are made widely available, with the results of such testing made publicly available to the extent possible. This evaluation should be done with the complete spectrum of model uses in mind, from deployment by model developers to distribution on platform/hosting intermediaries to specific business uses.

B. Support and Conduct Safety, Security, and Trustworthiness Research into Foundation Models and High Risk Models, including Downstream Uses

I. PERFORM INTERNAL GOVERNMENT RESEARCH

The U.S. government should engage in its own active research and analysis. The government should also continue to build capacity for a broad array of expertise and functions to conduct this research. Work being done by a variety of agencies in their respective areas of subject matter expertise could provide better insight into potential gaps that may need to be filled to promote an open ecosystem while addressing risks. For example, the U.S. Copyright Office is undergoing a comprehensive initiative to examine copyright issues raised by AI.[263] The Department of Energy's Frontiers in AI for Science, Security, and Technology (FASST)[264] initiative plans to leverage the departments' supercomputers to provide insights into dual-use foundation models and better assess potential risks. The outcome of initiatives such as these could create a better sense of the state of play in different fields (e.g., for the U.S. copyright system, a more comprehensive understanding of the interplay between the "fair use" doctrine and the use of copyrighted works without permission from the rights holder to train AI models). Consequently, any research and data gathering should, where appropriate, involve collaboration between relevant government agencies.

Research into foundation models should not just include technical aspects of the models. It should also cover areas of research such as the effects of AI on human actions, privacy, legal ramifications, and downstream effects of dual-use foundation models. This research should also address, for instance, the potential ability of these models to increase CBRN risks, in particular, bio risks, as well as cybersecurity concerns, and risks of human deception.

II. SUPPORT EXTERNAL RESEARCH

The federal government should support external research on the risks and benefits related to dual-use foundation models. Research into available technical and non-technical mitigations for risks arising from dual-use foundation models with widely available model weights is also important to prioritize. This could include research into model explainability/interpretability and other approaches identified by research communities. Support could take the form of direct research grants, including through the National AI Research Institutes, or it could be provided by prioritizing such research through compute resource support programs like the proposed NAIRR.

> **"Research into foundation models should not just include technical aspects of the models. It should also cover areas of research such as the effects of AI on human actions, privacy, legal ramifications, and downstream effects of dual-use foundation models."**

C. Develop and Maintain Risk Portfolios, Indicators, and Thresholds

The U.S. government should identify specific risks, and then, for each identified risk, maintain one or more risk indicators. These can be technical indicators, such as multi-modal capabilities or the ability of AI agents to manipulate the external environment, or measurements of confabulation or racial bias. They could also be societal indicators, such as the breadth of adoption of a particular AI system or the availability of certain physical materials which could be used in conjunction with AI to create a threat.

One important class of metrics for open-weight foundation models is leading indicators. These are indicators of the risks, benefits, and capabilities that open-weight foundation models will – but do not currently – possess. It is important that the government maintain robust leading indicators of model capabilities, because harms from open models are difficult to undo once the weights are released. While the existing capabilities of closed-weight models are one leading indicator of the future capabilities, risks, and benefits of open-weight foundation models, they are not the only ones. Tracking the relative rate of advances between open- and closed-weight models, for example by comparing their performance on complex tasks over time, would help identify when a given open-weight model is poised to catch up to or surpass the capabilities of an existing closed-weight model. By creating these metrics, the government can better prepare for future risks and take advantage of future benefits as these technologies continue to rapidly evolve.

To actively monitor the open-weight foundation model ecosystem, the federal government should maintain a portfolio of risk cases, including unlikely risks and sociotechnical risks, that might arise from future open foundation models. Each such risk should be accompanied by (i) one or more leading indicators of risk, which can be social and/or technological, (ii) thresholds for each indicator, and (iii) a set of potential policy responses that could mitigate the risk. Benefit indicators should also be taken into account when risk-benefit calculations are important. When the indicator(s) meet the threshold(s), the government should consider intervening with one or more policy responses. An example of this scenario is given in the Appendix.

"To actively monitor the open-weight foundation model ecosystem, the federal government should maintain a portfolio of risk cases, including unlikely risks and sociotechnical risks, that might arise from future open foundation models."

The choice of thresholds and potential policy responses should weigh current and predicted future technical capabilities, relevant legal considerations, and downstream impacts.

In establishing thresholds and conducting assessments, the government should recognize that the evidence base for restrictions on dual-use foundation models with widely available model weights is evolving. The benefits that such models produce should be fully considered in establishing those thresholds, as well as the legal and international enforcement challenges in implementing restrictions.[265] Additionally, consideration should be given to whether each risk is better addressed with interventions in downstream pathways through which those risks materialize rather than in the availability of model weights.[266]

"Additionally, consideration should be given to whether each risk is better addressed with interventions in downstream pathways through which those risks materialize rather than in the availability of model weights."

STEP 2

Evaluate Evidence

Using a broad evidence base and specific risk indicators, the federal government should assess whether the marginal risks from open-weight models in a particular sector or use case warrants government action. Specifically, the federal government should:

A. Assess the Difference in Capabilities, Limitations, and Information Content between Closed and Open Models

The government should assess and monitor the length of the "policymaking runway": the length of time between when leading closed models achieve new capabilities and when open-weight models achieve those same capabilities, along with a wider set of indicators including persistent limitations, and information about training data and information content associated with open weight models.

Once a capability appears in an open-weight model, it may be impossible to wholly remove that capability from the open-weight foundation model ecosystem. Therefore, restrictions on open-weight models can be most effective only before a particular capability is released in an open-weight model. Likewise, a rich understanding of limitations can help downstream integrators make informed choices when selecting open models.

Many factors may affect the policymaking runway, and its length will affect the speed with which policymakers will need to respond to changes in capabilities, limitations, and information content of open models available in the open model ecosystem.

B. Develop benchmarks and definitions for monitoring and action.

There are a range of factors that should be considered when developing monitoring benchmarks and definitions, not only those listed in the EO definition of dual-use foundation models. Numerical measures such as the number of floating-point operations used in training provide rough estimates of model capabilities, and can be used as a first step to distinguish models that deserve further scrutiny. But to properly calibrate monitoring and policy interventions to the appropriate models, the US government should developing benchmarks and definitions for model capabilities that incorporate other factors as well. One reason for this is that, while numerical measures like the number of parameters/weights or floating point operations per second (FLOPS) are often related to a model's technical capabilities,[267] advances in algorithms, architectures, processors, and the complexities posed by multi-modal models may gradually cause any numerical metric to become outdated. For instance, the Executive Order refers to "tens of billions of parameters" in the definition of dual-use foundation model. However, Meta's Llama 3 8B, which did not exist at the time the Executive Order was written and does not have enough parameters to meet this definition, outperforms LLama 2 70B,[268] which does meet the definition, on a number of benchmarks.[269] With computing capabilities increasing exponentially over time,[270] it is quite possible that personal computers will someday be able to train highly capable and generalizable models comparable to today's most advanced foundation models.[271]

Furthermore, the risks and benefits of AI arise in complicated social and technical ways, which depend on the type of information processed by the model and the potential set of use cases. Evo, a state-of-the-art AI biological design tool that can work with proteins, DNA, and RNA,[272] seems to fit most of the requirements for a dual-use foundation model.[273] However, some biological design tools currently only involve approximately hundreds of millions of parameters – far less than in the dual-use foundation model definition. Many text-to-image and text-to-video models do not require more than 10 billion parameters.[274] A giant model is not required to make a deepfake video – it can be done on a personal computer.[275]

In addition to the number of parameters, there are many other features that make AI models potentially powerful and which may be useful in establishing benchmarks and definitions for monitoring and action. Policymakers and researchers should take into consideration the following non-exhaustive list of factors. The relative importance of each factor will vary depending on the situation:

1. **Number of parameters**
2. **Computing resources required to train a model**
3. **Training data** – dataset size and quality, nature and confidentiality of the data, difficulty of reproducing the data.
4. **Model architecture and training methods**
5. **Versatility** – the types of tasks a model can perform
6. **Potential risks** – explicitly identified use cases that create specific harms
7. **Access and adoption** – the number of people, or-

ganizations, and systems that use or are affected by the model

8. **Emergence** – the ability of a model to perform tasks that it was not explicitly trained for

9. **Evaluated capabilities** – performance on particular tasks, including non-technical tasks such as AI-human interactions[276]

10. **Information modalities** – the types of information the model can process, such as image, text, genetic data[277], biometric data[278], real-world sensing[279] or combinations of multiple types.

C. Maintain and Bolster Cross-disciplinary Federal Government Capacity to Evaluate Evidence

Effective monitoring, assessment, and decision-making will require cross-disciplinary expertise and resources. The U.S. government should encourage and hire this type of talent. Technical specialists and access to AI models will be necessary to stay current on model capabilities. But social scientists will also be necessary to understand the economic and social effects of dual-use foundation models with widely available model weights. Legal experts, including privacy, First Amendment, copyright, foreign policy, as well as human and civil rights scholars, should be consulted on the legal and constitutional implications of intervening or failing to intervene. Domestic and international policy analysts will help navigate the complexities of government decision-making. The government has made significant strides in increasing the Federal AI workforce through the AI Talent Surge launched by EO 14110. The United States should continue that trend by hiring top talent across the fields that foster AI-related skills.[280] Particular care should be taken to maintain effective cross-agency collaboration because the impacts of dual-use foundation models do not fit neatly in any one category.

STEP 3

Act on Evaluations

Given the varied nature of risks that foundation models can and will pose, the government should maintain the ability to undertake interventions, which should be considered once the risk thresholds described above are crossed such that the marginal risks substantially outweigh the marginal benefit. These interventions include **restrictions on access to models** (including model weights) and **other risk mitigation measures** as appropriate to the specific context when restrictions on widely available model weights are not justified or legally permissible.

A. Model Access Restrictions

One broad category of such interventions involves restricting access to, or requiring pre-release model licensing for, certain classes of dual-use foundation models or systems. At one end of this category is complete restriction of a model from being publicly distributed, including model weights and API access. A less extreme step would involve restricting the open sharing of model weights and allowing public access only to hosted models. These restrictions would impose substantial burdens on the open-weight model ecosystem and should require significant evidence of risk. There are many different ways to implement a structured access program that restricts access to model weights,[281] where government could set guidelines "for what capabilities should be made available, in what form, and to whom."[282] The government could also mandate that intermediary AI platforms ensure that restricted weights are not available on their platforms or are only available in select instances. These restrictions could potentially be effectuated through existing statutory authorities (such as the Export Administration Act) or through Congressional action, though this Report does not consider questions of legal authority in detail.

Any consideration of the appropriate scope or nature of these restrictions would require legal and constitutional analysis.[283] Intellectual property considerations, which are not the principal focus of this Report, would also inform the question of whether, and how far, to restrict.

Importantly, the effects of AI and potential causes of AI risk are not bound to any single country, and the effectiveness of restrictions on the distribution of model weights depends in significant part on international alignment on the appropriate scope and nature of those restrictions. The federal government should prioritize international collaboration and engagement on its policy concerning the governance of dual-use foundation-models with widely available model weights.

> **"Given the varied nature of risks that foundation models can and will pose, the government should maintain the ability to undertake interventions, which should be considered once the risk thresholds described above are crossed such that the marginal risks substantially outweigh the marginal benefit."**

The United States should also retain the ability to promote certain types of openness in situations that have the potential to pose risk, but for which there is not enough information. This could include structured access for researchers,[284] further information gathering on the part of the U.S. government, or funding for specific risk research.

B. Other Risk Mitigation Measures

Because the risks and benefits of dual-use foundation models are not solely derived from the model itself, appropriate policy measures may not concern the model weights specifically, depending on the nature of the risks and benefits.[285] The government should maintain the ability to respond with a wide range of risk mitigations in accordance with its legal authority. The foundation model ecosystem has many components, and in many cases the most effective risk reduction will happen downstream of the model weights. It is important to note that several enforcement agencies have indicated that their authorities apply to the latest developments in AI technology, for example to address discrimination and bias.[286]

Whether and how regulations apply throughout the AI stack is beyond the scope of this Report, but it is an area worth exploring. These mitigations will likely depend on the specific risk. For example, in cases where dual-use foundation models with widely available model weights enable creation of dangerous physical objects, restrictions on physical materials may be warranted.

Firm data privacy protections should be developed and adapted as foundation models continue to interact with, and draw data from, progressively larger data sets, processed at higher velocities, that touch on more parts of Americans' lives. Other mitigation measures might include better content moderation on online platforms to limit illegal or abusive generated content, improved spear-phishing filters for emails, user interface designs to highlight possible misinformation and limited accessibility to CBRN datasets. Effective mitigations could also include making potentially impacted systems more robust and resilient to harmful effects of AI. This could include minimizing the reach of disinformation campaigns, and providing support resources for human victims of AI-generated harms. Ultimately, a combination of education, experience, research, and proactive efforts by model creators will likely be necessary to help mitigate a broad array of risks.

ADDITIONAL GOVERNMENT ACTION

While actively monitoring risks, the government should also support openness in ways that enhance its benefits. This should include incentivizing social, technical, economic, and policy research on how to ensure open foundation models promote human well-being. Government agencies may be able to use their authorities or subject matter expertise to promote an open ecosystem while addressing risks. Fiscal policy could also be used to support open foundation models, for instance through subsidies for open models. One promising subsidy-based approach is the NAIRR, which has embedded open source and open science principles into its workplan.

The U.S. government should also continue leading international diplomacy and norm-setting efforts around open foundation models. This should include engagement with a broad spectrum of international partners and fora to ensure the benefits of open artificial intelligence are shared, while limiting the ability of bad actors to cause harm. The U.S. government should also work with its allies to ensure that the uses of open-weight foundation models support the principles of democratic representation and freedom, rather than autocracy and oppression.

Conclusion

The current evidence base of the marginal risks and benefits of open-weight foundation models is not sufficient either to definitively conclude that restrictions on such open-weight models are warranted, or that restrictions will never be appropriate in the future. Accordingly, we recommend a three-part framework for the federal government to actively monitor the evidence base for the risks and benefits of dual-use foundation models with widely available model weights: collecting evidence about their capabilities, limitations, and information content, evaluating that evidence against thresholds of concern, and potentially acting upon those evaluations through appropriate policy measures. The government should also incentivize global and domestic research and innovation that harnesses the many benefits of open foundation models.

Appendix: Monitoring Template

This template is meant to show how the decision-making process might work, rather than suggest specific mitigation strategies and thresholds. Actual risk cases should be maintained by subject-matter experts who can collectively understand, monitor, and evaluate all details of a particular scenario. Notably, multiple government agencies with specific domains are monitoring AI-related risks using their own techniques and should be deferred to in those areas.

Risk: Foundation models increase the number of people with the potential capability to create a weapon and decrease team sizes and coordination costs required, thus increasing the chance that a domestic malicious actor creates and uses one.

In this risk scenario the availability of foundation models increases access for wider portions of the population, perhaps through the use of an LLM that can walk an individual through the steps required to create a weapon. This risk is distinct from the risk posed by scientifically sophisticated actors creating new weapons with increased potency. The discovery of a new weapon could also involve a model specifically developed to handle specialized knowledge (such as a biological design tool), which requires specialized expertise to use.

Collecting Evidence:

To create a weapon, an individual may need both specialized knowledge and appropriate materials. As model capabilities change, evaluators would need to gather and maintain information about the changing knowledge and material needs of actors seeking to create specific categories of weapons, which would require expertise in both science and machine learning. Evaluators may need to keep multiple risk profiles for different risks. Specific risk indicators might include, along with progressively less restrictive values of those indicators:

1. **What level of specialized knowledge is required to use the foundation model to create the desired weapon?**
   a. Specialized doctoral degree or higher
   b. Specialized master's degree
   c. Specialized bachelor's degree or hobbyist "home scientists"
   d. Average adult
2. **Where can an individual get the materials to make the desired weapon?**
   a. Specialty supplier, heavy regulation such as licensed sellers and buyers
   b. Specialty supplier, light regulation such as purchase tracking
   c. Specialty supplier, no legal restrictions but typically has administrative barriers
   d. Local store or Internet search

To gather this information, evaluators could begin by red-teaming open and closed cutting-edge models. Subject-matter experts would consider the additional assistance that the model provides in creation of a weapon. They would also consider the equipment required, including whether the model finds methods to use more easily available materials than might be purchased through a laboratory supplier. Other methods might be used, as determined by the subject-matter experts.

Evaluating Evidence and Acting on Evaluations:

The grid below shows a set of potential mitigation options which are dependent on the risk indicators. The government agency responsible for managing the risk scenario would choose from potential mitigation options, which could involve technical restrictions or a variety of other non-model-oriented actions designed to reduce risk. Developing such a matrix would require an understanding of different legal and regulatory authorities and may involve collaboration between agencies. In the example decision matrix below, the value in each entry shows possible mitigation options, which the agency may or may not decide to recommend.

Mitigation Options:

0. **Do nothing**
1. **Restrict open model weights & access to closed models, for specific classes of models**
2. **Restrict access to specific materials**
3. **Security controls on API-based fine-tuning of closed models using specific types of data (biological, chemical, etc.)**

Who is enabled by AI to create a weapon?					
Where can an individual get the materials to make a weapon?		Average person	Specialized bachelor’s degree / hobbyist	Specialized master’s degree	Specialized doctoral degree
	Local store/ Internet search	1 or 2 or 3	1 or 2 or 3	1 or 2 or 3	2 or 3
	Specialty supplier, no legal restrictions	1 or 2 or 3	1 or 2 or 3	1 or 2	2
	Specialty supplier, light legal burden	1 or 2 or 3	1 or 2 or 3	1 or 2	0
	Specialty supplier, heavily regulated	1	1	0	0

Example: If an individual with a specialized master’s degree can use an LLM to make a weapon with materials from a specialty supplier with no legal burden, then the government should consider either (1) restricting access to specific classes of models/weights or (2) restricting access to specific materials.

Endnotes

[1] Exec. Order No. 14,110 (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[2] We define limited access as AI models that do not give access to model weights, source code, or training data.

[3] All mentions of specific companies or services in this report are referential and not intended to imply normativity, either positive or negative, regarding the company or service.

[4] Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights, 89 Fed. Reg. 14059 Pub (Feb. 26, 2024).

[5] Exec. Order No. 14,110 (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[6] Exec. Order No. 14,110 (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[7] This provision of Executive Order 14110 refers to the as-yet speculative risk that AI systems will evade human control, for instance through deception, obfuscation, or self-replication. Harms from loss of control would likely require AI systems which have capabilities beyond those known in current systems and have access to a broader range of permissions and resources than current AI systems are given. However, open models introduce unique considerations in these risk calculations, as actors can remove superficial safeguards that prevent model misuse. They can also customize, experiment with, and deploy models with more permissions and in different contexts than the developer originally intended. Currently, AI agents who can interact with the world lack capabilities to independently perform complex or open-ended tasks, which limits their potential to create loss of control harms. Hague, D. (2024). Multimodality, Tool Use, and Autonomous Agents: Large Language Models Explained, Part 3. Center for Security and Emerging Technology. https://cset.georgetown.edu/article/multimodality-tool-use-and-autonomous-agents/ ("While LLM agents have been successful in playing Minecraft and interacting in virtual worlds, they have largely not been reliable enough to deploy in real-life use cases. . . Today, research often focuses on getting autonomous LLMs to perform specific, defined tasks like booking flights."). Developing capable AI agents remains an active research goal in the AI community. Xi, Z., et al. (2023). The Rise and Potential of Large Language Model Based Agents: A Survey. ArXiv.org. https://doi.org/10.48550/arXiv.2309.07864. However, given the nascent stage of these efforts, this Report cannot yet meaningfully discuss these risks in greater depth.

[8] Exec. Order No. 14,110 (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[9] Bommasani, R., Hudson, D.A., Adeli, E. et al (2021). On the opportunities and risks of foundation models. ArXiv preprint. https://arxiv.org/abs/2108.07258.

[10] Exec. Order No. 14,110 (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[11] The Report defines AI models as "a component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs," as defined in the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" Executive Order. (Exec. Order No. 14,110 (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/).

[12] Yuksel, S., et al., (2012). Twenty Years of Mixture of Experts. IEEE Transactions on Neural Networks and Learning Systems, 23(8), 1177–1193; Vaswani, A. et al., (2017). Attention is All You Need. https://arxiv.org/abs/1706.03762 ; What Is RLHF? (n.d.). aws.amazon.com. https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/ ; Whang, O. (2024, April 30). From Baby Talk to Baby A.I. NYTimes. https://www.nytimes.com/2024/04/30/science/ai-infants-language-learning.html.

[13] Lin, Y.-T., & Chen, Y.-N. (2023). Taiwan LLM: Bridging the Linguistic Divide with a Culturally Aligned Language Model. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.17487.

[14] Representing weights with lower-precision numbers. See, e.g., Hugging Face. Quantization. https://huggingface.co/docs/optimum/en/concept_guides/quantization.

[15] Various methods that end up removing parameters from an AI model. See, e.g., Pruning Tutorial. PyTorch. https://pytorch.org/tutorials/intermediate/pruning_tutorial.html.

[16] Criddle, C. & Madhumita M. (2024, May 8). Artificial intelligence companies seek big profits from 'small' language models. Financial Times. https://www.ft.com/content/359a5a31-1ab9-41ea-83aa-5b27d9b24ef9.

[17] A CNAS report found that, if trends continue, frontier AI training could require 1,000 times more compute power than GPT-4 by the late 2020s/early 2030s, and training costs for leading models double approximately every 10 months. However, note that the amount of compute power an actor saves depends on the amount on inference they need to perform on the model. Future-Proofing Frontier AI Regulation. (2024, March 13). https://www.cnas.org/publications/reports/future-proofing-frontier-ai-regulation.

[18] See e.g.,Vincent, J. (2023, March 8). Meta's powerful AI language model has leaked online — what happens now? The Verge. https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse ; Hubinger, E., Denison, et al. (2024, January 17). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. ArXiv. https://doi.org/10.48550/arXiv.2401.05566.

[19] Zhan, Q. et al., (2023). Removing RLHF Protections in GPT-4 via Fine-Tuning. UIUC, Stanford. https://openreview.net/pdf?id=NyGm96pC3n.

[20] See, e.g., Partnership on AI Comment at 6, ("Some of the features of open models that may be relevant to assessing differential risk include that open release of model weights is irreversible, and that moderation/monitoring of open models post-release is challenging.").

[21] See, e.g., Hugging Face Comment at 3 ("Model weights can be shared individually between parties, on platforms with or without documentation and with or without access management, and via p2p/torrent.").

[22] See Goldman, S. (2023, December 8). Mistral AI bucks release trend by dropping torrent link to new open source LLM. VentureBeat. https://venturebeat.com/ai/mistral-ai-bucks-release-trend-by-dropping-torrent-link-to-new-open-source-llm/; Coldewey, D. (2023, September 27). Mistral AI makes its first large language model free for everyone. TechCrunch. https://techcrunch.com/2023/09/27/mistral-ai-makes-its-first-large-language-model-free-for-everyone/. However, note that not all AI models receive such attention when released. See GitHub Comment at 3 ("Wide availability of model weights is a function of discovery, governed by online platforms. Even for content posted publicly on the internet, the default state is obscurity. Whether content is widely available will depend on ecosystem activity, distribution channels, and, particularly, sharing on platforms that enable virality. Ecosystem monitoring and governance can help inform and implement risk-based mitigations for widely available model weights.").

[23] See, e.g., Holistic AI Comment at 6 ("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."); Future of Life Institute Comment at 7 ("AI systems claimed to be 'open' lie across a spectrum of access, each carrying different levels of benefits and risks.") (quotation marks in original); OTI Comment at 4 ("There is no easy binary that opposes 'open' and 'closed' AI models. Commentary and research that suggest otherwise unhelpfully distort the reality—which AI technical and governance experts have repeatedly explained—that AI models sit somewhere on a spectrum or 'gradient' of openness.") (internal citation omitted) (quotation mark in original); SIIA Comment at 1 ("First, openness should be viewed across a gradient, with model weights as one component of an AI system that can be made available to third parties in varying degrees."); Intel Comments at 4 ("Release of foundation models present a gradient of openness options.") (internal citation omitted); CCIA Comment at 2 ("[T]here is a broad spectrum based on how much of the system is open and how that system is managed."); Google Comment at 4 ("Assuming 'openness' in AI as a binary choice between open and closed approaches fails to capture important nuances. Rather, it is important to think about 'open' and 'widely available' as existing on a gradient of access, which offers a better conceptual frame.") (internal footnote omitted) (quotation marks and italics in original); Michael Weinberg Comment at 4 ("The definition of 'open' in the context of foundational models continues to evolve, with many different approaches making a reasonable claim to being open depending on the context and intended use.") (internal citation omitted) (quotation marks in original). See also generally Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations. Hugging Face. https://arxiv.org/pdf/2302.04844 (discussing gradients of AI system release); Bommasani, R., et al., (2023). Considerations for Governing Open Foundation Models. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf (similar).

[24] Shevlane, T. (2022). Structured Access: An Emerging Paradigm for Safe AI Deployment. University of Oxford. https://arxiv.org/abs/2201.05159.

[25] Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations. Hugging Face. https://arxiv.org/pdf/2302.04844.

[26] Contractor, D., et al. (2022, June). Behavioral use licensing for responsible AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 778-788).

[27] Mitchell, M., et al. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

[28] See, e.g., AI Accountability Policy Report, National Telecommunications and Information Administration. (2024, March). https://www.ntia.gov/sites/default/files/2024-04/ntia-ai-report-print.pdf at 28 (noting that datasheets "provide salient information about the data on which the AI model was trained, including the 'motivation, composition, collection process, [and] recommended uses' of the dataset").

[29] Exec. Order No. 14,110 (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[30] See, e.g., AI Policy and Governance Working Group Comment at 2 ("The federal government should prioritize understanding of marginal risk. The risks of open foundation models do not exist in a vacuum. To properly assess the risks of open foundation models, and whether regulations should single out open foundation models, the federal government should directly compare the risk profile to those of closed foundation models and existing technologies. In its report, the NTIA should foreground the marginal risk of open foundation models by directing government agencies to conduct marginal risk assessments, fund marginal risk assessment research, and incorporate marginal risk assessment into procurement processes."); Rishi Bommasani et al. at 1 ("Foundation models present tremendous benefits and risks to society as central artifacts in the AI ecosystem. In addressing dual use foundation models with widely available weights, the National Telecommunications and Information Administration (NTIA) should consider the marginal risk of open foundation models, defined as the extent to which they increase risk relative to closed foundation models or preexisting technologies like search engines.") (internal footnote omitted); CTA Comment at 6 ("NTIA's consideration of risks associated with open weight models should focus on marginal

risks arising from such models."); Public Knowledge Comment at 3 ("The conversation around open foundation models is significantly enriched by a nuanced understanding of the marginal risks they pose compared to their closed counterparts and existing technologies."); OTI Comment at 17 ("NTIA and other U.S. government agencies must focus vague discussions about the risks of open AI models on the study and precise articulation of the marginal risk these models pose."); Holistic AI Comment at 10 ("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") (internal hyperlink omitted); CDT Comment at 14 ("In evaluating the risks of [open foundation models], we must consider them in comparison to the existing risks enabled by closed models, by access to existing technologies such as the internet, and by smaller models that carry similar risks but for which controlling proliferation would be much harder if not impossible. In other words, we must consider the marginal risk of [open foundation models].") (italics in original) (internal citation omitted); Mozilla Comment at 13 ("Debates around safety and 'open source' AI should center marginal risk[.]") (quotation marks in original); Microsoft Comment at 1-2 ("We recommend [. . .] [p]romoting risk and impact assessments that are grounded in the specific attributes of widely available model weights that present risk, the marginal risk of such availability compared to existing systems[.] [. . .]"); PAI Comment at 6 ("In assessing the risk posed by open foundation models, and appropriate measures to address those risks, policy makers should focus on the marginal risks associated with open access release.") (internal hyperlink omitted); Databricks Comment at 2 ("The benefits of open models substantially outweigh the marginal risks, so open weights should be allowed, even at the frontier level[.]"); Meta Comment at 16 ("In order to precisely identify and assess risks uniquely presented by open foundation models, it is important to apply a 'marginal risk analysis' that takes account of the risks of open models compared to: (1) preexisting technologies, and (2) closed models.") (internal citation omitted); (quotation marks in original) GitHub Comment at 3 ("Evidence of harmful capabilities in widely available model weights and their use should consider baselines of closed, proprietary AI capabilities and the availability of potentially dangerous information in books and via internet search. [. . .] Today, available evidence of the marginal risks of open release does not substantiate government restrictions."); BSA Comment at 3 ("Any specific policy options for open foundation models should be considered only as any marginal risk posed by such models are better understood."); U.S. Chamber of Commerce at 3 ("As indicated in the NIST [Risk Management Framework] 1.0, 'Risk tolerance and the level of risk acceptable to organizations or society are highly contextual and application and use-case specific.' This is why we believe it is essential for NTIA to focus on the marginal risk, which is context-specific.") (internal citation omitted) (quotation marks in original); AI Healthcare Working Group at 1 ("The risks of technology are real, but their promise outweighs those risk and those risks should be viewed and evaluated in the context of marginal risk."); Johns Hopkins Center for Health Security Comment at 5 ("As Sayesh Kapoor and colleagues caution, it is important to consider the marginal risk that open models pose above preexisting technologies.") (internal citation omitted). See also generally Center for Democracy & Technology, & et al. (March, 25, 2024). RE: Openness and Transparency in AI Provide Significant Benefits for Society. https://cdt.org/wp-content/uploads/2024/03/Civil-Society-Letter-on-Openness-for-NTIA-Process-March-25-2024.pdf (letter from civil society organizations promoting a marginal risk assessment); Kapoor S. et al., (2024). On the Societal Impact of Open Foundation Models. ArXiv. https://arxiv.org/pdf/2403.07918 (presenting a marginal risk framework).

[31] See Executive Order 14110, section 4.6.

[32] See, e.g., Center for AI Policy Comment at 6 ("We find that the timeframe between closed and open models right now is around 1.5 years. We can arrive at this conclusion by analyzing benchmark performance between current leading open weight AI models and the best closed source AI models."); Unlearn.AI Comment at 2 ("Estimating the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks is possible by looking at the gaps in human-evaluated performance between open foundation models and closed counterparts. While this is highly dependent on the specific AI model and its application domain, we can look towards a few examples. At the moment, it takes about 6 months to 1 year for similarly performing open models to be successfully deployed after the deployment of OpenAI's closed models. The time gap between proprietary image recognition models and high-quality open-source alternatives has narrowed relatively quickly due to robust community engagement and significant public interest. In contrast, more niche or complex applications, such as those requiring extensive domain-specific knowledge or data, might see longer timeframes before competitive open models emerge."); Databricks Comment at 3 ("Databricks believes that major open source model developers are not far behind the closed model developers in creating equally high performance models, and that the gap between the respective development cycles may be closing.") (internal citation omitted); Stability AI Comment at 17 ("There is ample evidence that closed models exhibiting category state of the art performance will be matched by open models in due course. Previously, it took ~28 months before an open model such as GPT-J from EleutherAI approached the performance of a closed model such as GPT-2 from Open AI on common benchmarks. That gap is closing. Only ~eight months elapsed before open models such as Llama 2-70B from Meta rivaled GPT-3.5 from Open AI, and only ~ten months elapsed before Falcon-180B from the Technology Innovation Institute (funded by the Abu Dhabi Government) exceeded GPT-3.5 performance.") (internal citation omitted). But see Meta Comment at 13 ("It is not possible to generally estimate this timeframe given the variables involved, including the model deployment developers' business models and whether, in the case of Llama 2, they download the model weights from Meta directly or accessed it through third-party services like Azure or AWS."); Hugging Face Comment at 3 ("Timelines vary and estimates will change by utility of the model and costs."); EleutherAI Comment at 21 ("The timeframe between deployment of an open and closed equally-performing model is difficult to predict reliably. The primary blocker for the capabilities of open models is funding, which can disappear at the whim of a handful of well-resourced individuals. [. . .]" See also CSET Comment at 2 ("The best way to gauge such timeframes may be to directly contact organizations designing foundation models and acquire information regarding their model performance and release strategies. This is the most viable way to get these estimations, although these organizations may not have the will or obligation to provide such information.").

[33] Fine-tuning away Llama 2-Chat 13B's safety features while retaining model performance costs less than $200. See, Gade, P., et al. (2023, October 31). BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B. ArXiv. https://doi.org/10.48550/arXiv.2311.00117.

[34] Filters applied to generated content that prevent prohibited material from being returned to the user.

[35] Lists of words, phrases, and topics that cannot be generated.

[36] Measures intended to prohibit prompts that attempt to circumvent the aforementioned safety features. However, see generally, Can Foundation Models Be Safe When Adversaries Can Customize Them? (2023, November 2). Hai.stanford.edu. https://hai.stanford.edu/news/can-foundation-models-be-safe-when-adversaries-can-customize-them; Henderson, P., et al. (2023, August 8). Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models. ArXiv. https://doi.org/10.48550/arXiv.2211.14946.

[37] See, Seger, E., et al. (2023, October 9). Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. Social Science Research Network. https://doi.org/10.2139/ssrn.4596436; Boulanger, A. (2005). Open-source versus proprietary software: Is one more reliable and secure than the other? IBM Systems Journal, 44(2), 239–248. https://doi.org/10.1147/sj.442.0239; Gade, P., et al. (2023, October 31). BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B. ArXiv. https://doi.org/10.48550/arXiv.2311.00117.

[38] Mouton, Christopher A., et al., The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. Santa Monica, CA: RAND Corporation, 2024. https://www.rand.org/pubs/research_reports/RRA2977-2.html. See also CDT Comment at 19.

[39] Zhan, Q. et al., (2023). Removing RLHF Protections in GPT-4 via Fine-Tuning. UIUC, Stanford. https://openreview.net/pdf?id=NyGm96pC3n.

[40] See, Seger, E., et al. (2023, October 9). Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. Social Science Research Network. https://doi.org/10.2139/ssrn.4596436.

[41] For instance, a method discovered to jailbreak Meta's Llama 2 works on other LLMs, such as GPT-4 and Claude. Seger, E., et al. (2023). Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. Social Science Research Network. https://doi.org/10.2139/ssrn.4596436.

[42] Li, N., et al. (2024, May 15). The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. ArXiv. https://doi.org/10.48550/arXiv.2403.03218; Lynch, A., et al. (2024, February 26). Eight Methods to Evaluate Robust Unlearning in LLMs. ArXiv. https://doi.org/10.48550/arXiv.2402.16835.

[43] Mouton, C. A., et al. (2024, January 25). The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

[44] Mouton, C. A., et al. (2024, January 25). The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

[45] Congressional Research Services (2023, November 23) Artificial Intelligence in the Biological Sciences: Uses, Safety, Security, and Oversight, https://crsreports.congress.gov/product/pdf/R/R47849.

[46] See Johns Hopkins Center for Health Security Comment at 5. See Johns Hopkins Center for Health Security Comment at 5 ("Indeed, less than a month after Evo was released, it had already been fine-tuned on a dataset of adeno-associated virus capsids, ie, protein shells used by a class of viruses that infect humans. As this case suggests, when a model's weights are publicly available, a developer's decision not to endow the model with dangerous capabilities is far from final.").

[47] See generally, Nguyen, E., et al. (2024, February 27). Evo: DNA foundation modeling from molecular to genome scale. Arc Institute. https://arcinstitute.org/news/blog/evo; Abramson, J., et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature (2024). https://doi.org/10.1038/s41586-024-07487-w.

[48] Sandbrink, J. (2023). Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. ArXiv. https://arxiv.org/abs/2306.13952.

[49] One counterpoint is Google Alpha fold and predicting protein folding. Putting the power of AlphaFold into the world's hands. (2024, May 14). Google DeepMind. https://deepmind.google/discover/blog/putting-the-power-of-alphafold-into-the-worlds-hands/.

[50] See EPIC Comment at 4 fn 15. See Liwei Song & Prateek Mittal, Systematic Evaluation of Privacy Risks of Machine Learning Models, 30 Proc. USENIX Sec. Symp. 2615, 2615 (2021). Hurdles to unlearning data are at the core of recent FTC cases requiring AI model deletion. See Jevan Hutson & Ben Winters, America's Next 'Stop Model!': Model Deletion, 8 Geo. L. Tech. Rev. 125, 128–134 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4225003.

[51] Mouton, C. A., et al. (2024, January 25). The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

[52] Terwilliger, T. C., et al. (2023). AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nature Methods, 1–7. https://doi.org/10.1038/s41592-023-02087-4.

[53] Fang, R., et al. (2024). LLM Agents can Autonomously Exploit One-day Vulnerabilities. https://arxiv.org/html/2404.08144v1.

[54] State of Open Source AI Book 2023 Edition. (2024). https://book.premai.io/state-of-open-source-ai/#contributing.

[55] Cyber attackers could use foundation models in assisting in the design or deployment of sophisticated malware, including viruses, ransomware, and

Trojans. For instance, Llama 2, a foundation model with widely available model weights developed by Meta, has already helped cyber-attackers design tools to illicitly download employees' login information. Ray, T. (2024, February 21). Cybercriminals are using Meta's Llama 2 AI, according to CrowdStrike. Zdnet, https://www.zdnet.com/article/cybercriminals-are-using-metas-llama-2-ai-according-to-crowdstrike/. Initial evidence suggests that some closed-weight foundation models can be used to "autonomously hack websites, performing tasks as complex as blind database schema extraction and SQL injections without human feedback" and to "autonomously find[] [cybersecurity] vulnerabilities in websites in the wild." Fang, R., et al. (2024, February 15). LLM Agents can Autonomously Hack Websites. ArXiv.org. https://doi.org/10.48550/arXiv.2402.06664. The National Cyber Security Centre of the Government of the United Kingdom assesses that "in the near term, [vulnerability detection and exploitation] will continue to rely on human expertise, meaning that any limited uplift [in cyberattack threat] will highly likely be restricted to existing threat actors that are already capable. . . . However, it is a realistic possibility that [constraints on expertise, equipment, time, and financial resourcing] may become less important over time, as more sophisticated AI models proliferate and uptake increases." National Cyber Security Centre. (2024, January 24). The near-term impact of AI on the cyber threat. www.ncsc.gov.uk. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat. Should these attacks successfully target electrical grids, financial infrastructures, government agencies, and other entities critical to public safety and national security, the security implications could be significant.

[56] A virus replicates itself by modifying other programs and inserting its code into those programs.

[57] Ransomware is malware that holds a device or data hostage until the victim pays a ransom to the hacker.

[58] A Trojan malware attack misleads users by disguising itself as a standard program.

[59] Cybercriminals are using Meta's Llama 2 AI, according to CrowdStrike. (n.d.). ZDNET. https://www.zdnet.com/article/cybercriminals-are-using-metas-llama-2-ai-according-to-crowdstrike.

[60] Cyber attackers could possibly use dual-use foundation models with widely available model weights to perform cyberattacks on closed models or extract data from them. Actors could (i) poison models' training data with an influx of synthetically generated content, (ii) steal model weights and other proprietary model infrastructure content through generated "jailbreaking" prompts, and (iii) leverage open models to access individual data from closed models trained on private data, which introduces privacy and autonomy concerns. See Nasr, M. (2023). Scalable Extraction of Training Data from (Production) Language Models. Google DeepMind, University of Washington, Cornell, CMU, UC Berkeley, and ETH Zurich.

[61] See National Cyber Security Centre. (2024, January 24). The near-term impact of AI on the cyber threat. www.ncsc.gov.uk. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat ("The impact of AI on the cyber threat is uneven; both in terms of its use by cyber threat actors and in terms of uplift in capability.").

[62] See National Cyber Security Centre. (2024, January 24). The near-term impact of AI on the cyber threat. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat at 5-7.

[63] MITRE ATT&CK. (n.d.). https://attack.mitre.org/; Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. ArXiv. https://arxiv.org/pdf/2403.07918.

[64] Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. ArXiv. https://arxiv.org/pdf/2403.07918.

[65] U.S. Cybersecurity and Infrastructure Security Agency. May 2024. CISA Response to NTIA Request for Information on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights ("Foundational models have at least two classes of potential harms. […] The second class involves impacts that are undesired by those deploying the models (e.g., cybersecurity vulnerability in a model deployed by a critical infrastructure entity). […] Creators and deployers of open foundation models can take steps to mitigate the second class of harms by using a "safe by design" approach and building in protections to their model. This may address cybersecurity vulnerabilities or other forms of harms such as biases. Responsibly developed open foundation models are likely to be less susceptible to harms and misuse, on the whole, than models that cannot be publicly audited.").

[66] U.S. Cybersecurity and Infrastructure Security Agency. May 2024. CISA Response to NTIA Request for Information on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights ("Foundational models have at least two classes of potential harms. […] The second class involves impacts that are undesired by those deploying the models (e.g., cybersecurity vulnerability in a model deployed by a critical infrastructure entity). […] Creators and deployers of open foundation models can take steps to mitigate the second class of harms by using a "safe by design" approach and building in protections to their model. This may address cybersecurity vulnerabilities or other forms of harms such as biases. Responsibly developed open foundation models are likely to be less susceptible to harms and misuse, on the whole, than models that cannot be publicly audited.").

[67] CISA. (n.d.) Secure by Design. https://www.cisa.gov/securebydesign.

[68] While Security-BERT is not a dual-use foundation model because it does not have "at least tens of billions of parameters," as required by Section 3(k) of Executive Order 14110, its capabilities may be indicative of the capabilities of dual-use foundation models.

[69] Ferrag, M. et al., (2024). Revolutionizing Cyber Threat Detection with Large Language Models: A privacy-preserving BERT-based Lightweight Model for IoT/IIoT Devices. Technology Innovation Institute.

[70] Alam, M. (2023). Recasting Self-Attention with Holographic Reduced Representations. ArXiv. https://arxiv.org/abs/2305.19534. Deng, Y. (2022). Large Language Models are Zero-Shot Fuzzers: Fuzzing Deep-Learning Libraries via Large Language Models. ArXiv. https://arxiv.org/abs/2212.14834.

[71] Rotlevi, S. (2024, February 15). AI Security Tools: The Open-Source Toolkit. https://www.wiz.io/academy/ai-security-tools; Hughes, C. (2024, January 16).

The OWASP AI Exchange: An open-source cybersecurity guide to AI components. https://www.csoonline.com/article/1290876/the-owasp-ai-exchange-an-open-source-cybersecurity-guide-to-ai-components.html.

72 Vulnerability detection and scanning is an important tool in the cybersecurity defense toolbox, as is well demonstrated by the Cybersecurity and Infrastructure Security Agency's focus on vulnerability scanning in its "Cyber Hygiene" services, which are designed to improve the cybersecurity of government and critical infrastructure organizations. See CISA. (n.d.). Cyber Hygiene Services. https://www.cisa.gov/cyber-hygiene-services. See also The White House. (2023). National Cybersecurity Strategy. https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf at 21 (describing "new tools for secure software development, software transparency, and vulnerability discovery" as important factors in defining cybersecurity obligations for software developers).

73 Zellers, R. et al., (2020). Defending Against Neural Fake News. ArXiv preprint. https://arxiv.org/pdf/1905.12616; Kirchenbauer, J. (2024); On the Reliability of Watermarks for Large Language Models. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2306.04634; Liu, H., et al. (2023). Chain of Hindsight Aligns Language Models with Feedback. ArXiv. https://doi.org/10.48550/arXiv.2302.02676; Belrose, N. LEACE: Perfect linear concept erasure in closed form. 37th Conference on Neural Information Processing Systems (NeurIPS 2023). https://proceedings.neurips.cc/paper_files/paper/2023/file/d066d21c619d0a78c5b557fa3291a8f4-Paper-Conference.pdf; Bhardwaj, R. & Poria, S. (2023). Red-Teaming Large Language Models using Chain of Utterances for Safety-Alignment. ArXiv preprint. https://doi.org/10.48550/arXiv.2308.09662; Zou, A. et al (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. ArXiv preprint. https://doi.org/10.48550/arXiv.2307.15043.

74 See CCIA Comment at 1-2 ("Open models also present advantages in AI governance, being easier to understand and test.").

75 Mozilla Open Source Audit Tooling (OAT) Project. (n.d.). Mozilla. https://foundation.mozilla.org/en/what-we-fund/oat/.

76 Lambert, N. et al. (2024, June 8). RewardBench: Evaluating Reward Models for Language Modeling. ArXiv. https://doi.org/10.48550/arXiv.2403.13787.

77 Hubinger, E., et al. (2024, January 17). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. ArXiv. https://doi.org/10.48550/arXiv.2401.05566; Measuring the impact of post-training enhancements. (n.d.). METR's Autonomy Evaluation Resources. https://metr.github.io/autonomy-evals-guide/elicitation-gap/.

78 See also AI Accountability Policy Report, National Telecommunications and Information Administration. (2024, March). https://www.ntia.gov/sites/default/files/2024-04/ntia-ai-report-print.pdf at 70 (noting that "[i]ndependent AI audits and evaluations are central to any accountability structure []").

79 See Anthony Barret Comment at 2 ("Although both closed and open models can pose some such risks, unsecured models pose unique risks in that safety and ethical safeguards that were implemented by developers can be removed relatively easily from models with widely available weights (e.g., via fine tuning).") (citation omitted).

80 See generally, Hofmann, et al. (2024, March 1). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. ArXiv. https://doi.org/10.48550/arXiv.2403.00742.

81 Examining Malicious Hugging Face ML Models with Silent Backdoor. (2024, February 27). JFrog. https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/.

82 Nestor Maslej, et al. Artificial Intelligence Index Report 2024. (2024 April). https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf.

83 Markus Anderljung et al. (2023, July 6). Frontier AI Regulation: Managing Emerging Risks to Public Safety. ArXiv. https://doi.org/10.48550/arXiv.2307.03718.

84 Center for a New American Security Comment at 9 ("[L]eading Chinese labs have applied the architecture and training process of Meta's Llama models to train their own models with similar levels of performance.").

85 Szabadföldi, István. (2021). Artificial Intelligence in Military Application – Opportunities and Challenges. Land Forces Academy Review. 26. 157-165. 10.2478/raft-2021-0022.

86 Mozur, P. et al. (2024, February 21). China's Rush to Dominate A.I. Comes With a Twist: It Depends on U.S. Technology. NYTimes. https://www.nytimes.com/2024/02/21/technology/china-united-states-artificial-intelligence.html.

87 See Transformative Futures Institute comment at 2 ("To maximize the benefit of foundation models, developers can offer structured access to models (e.g., through APIs) to a wide range of users while maintaining the capability to prevent harmful misuse, block or filter dangerous content, and conduct continual safety evaluations.").

88 See OpenAI Comment at 2 (" For example, we recently partnered with Microsoft to detect, study, and disrupt the operations of a number of nation-state cyber threat actors who were abusing our GPT-3.5-Turbo and GPT-4 models to assist in cyber-offensive operations. Disrupting these threat actors would not have been possible if the weights of these at-the-time leading models had been released widely, as the same cyber threat actors could have hosted the model on their own hardware, never interacting with the original developer.").

89 See RAND comment at 4 ("Coordination with foreign partners will be critical to manage risks from open foundation models. Attempts to control any open foundation models should consider which foreign actors could re-create equally capable models, as those operating in") ; See Connected Health Initiative at 4 ("Support international harmonization. We urge NTIA to maintain a priority for supporting risk-based approaches to AI governance in markets abroad

and through bilateral and multilateral agreements.").

90 Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. at 2. ArXiv. https://arxiv.org/pdf/2403.07918.

91 Nestor Maslej, et al. Artificial Intelligence Index Report 2024. (2024 April). https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf.

92 See, e.g., Johns Hopkins Center for Health Security Comments at 2 (Acknowledging that at current, Evo, a biological design tool (BDT) "…is a 7-billion parameter model and so [is] below the EO size threshold for a dual-use foundation model…"); and The Future Society Comments at fn. 20, citing Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. ArXiv. https://arxiv.org/pdf/2403.07918. ("AI-generated pornography based on Stable Diffusion offshoots quickly spread across the internet, including images resembling real people generated without their consent."). The largest version of Stable Diffusion, Stable Diffusion XL, has only 3B parameters (HuggingFace).

93 Ecosystem Graphs for Foundation Models. (n.d.). https://crfm.stanford.edu/ecosystem-graphs/index.html?mode=table.

94 See generally, Foundation Model Privacy. IBM Research. https://research.ibm.com/projects/foundation-model-privacy ("Language models have an inherent tendency to memorize and even reproduce in their outputs text sequences learned during training, may this be pre-training, fine-tuning or even prompt-tuning. If this training data contained sensitive or personal information, this could result in a major privacy breach."); Hartmann, V., et al. (2023). SoK: Memorization in General-Purpose Large Language Models. https://arxiv.org/pdf/2310.18362 ("In many cases of interest, such as personal identifiers, social security numbers or long passages of verbatim text, it is unlikely that a model could hallucinate the target information or gain knowledge of it through reasoning."); Huang, J., Shao, H., & Chang, K. C.-C. (2022, May 25). Are Large Pre-Trained Language Models Leaking Your Personal Information? ArXiv. https://arxiv.org/abs/2205.12628 ("We find that PLMs do leak personal information due to memorization.").

95 See Thorn Comments at 1 ("One concrete risk that is already manifesting as a harm occurring today, is the misuse of broadly shared and open source foundation models to make AI-generation child sexual abuse material. This technology is used to newly victimize children, as bad actors can now easily sexualize benign imagery of a child to scale their sexual extortion efforts… . This technology is further used in bullying scenarios, where sexually explicit AI-generated imagery is being used by children to bully and harass others.") (citations omitted).

96 2023 State of Deep Fakes. (2023). https://www.homesecurityheroes.com/state-of-deepfakes/.

97 Home Security Heroes. (2023). 2023 State of Deepfakes: Realities, Threats, and Impacts. Home Security Heroes. https://www.homesecurityheroes.com/state-of-deepfakes/ ("99% of the individuals targeted in deepfake pornography are women."), and Eaton, A. A., Ramjee, D., & Saunders, J. F. (2023). The Relationship between Sextortion during COVID-19 and Pre-pandemic Intimate Partner Violence: A Large Study of Victimization among Diverse U.S Men and Women. Victims & Offenders, 18(2), 338–355. https://doi.org/10.1080/15564886.2021.2022057.

98 Internet Watch Foundation (2023). How AI is being abused to create child sexual abuse imagery. https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf ; Kang, C. (2024). A.I.-Generated Child Sexual Abuse Material May Overwhelm Tip Line. NYTimes. https://www.nytimes.com/2024/04/22/technology/ai-csam-cybertipline.html.

99 How AI is being abused to create child sexual abuse imagery. (2023). Internet Watch Foundation. https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.

100 Kapoor S. et al., (2024). On the Societal Impact of Open Foundation Models. ArXiv. https://arxiv.org/pdf/2403.07918.

101 Thiel, D., Stroebel, M., & Portnoff, R. (2023, June 24). Generative ML and CSAM: Implications and mitigations. FSI. https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations.

102 How AI is being abused to create child sexual abuse imagery. (2023). Internet Watch Foundation. https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.

103 Thiel, D. (2023). Generative ML and CSAM: Implications and Mitigations. Stanford. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf.

104 Stable Diffusion Public Release. (2023, August 22). Stability AI. https://stability.ai/news/stable-diffusion-public-release.

105 Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Internet Observatory. https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf.

106 Open Technology Institute Comments at 11, citing Thiel, D. (2023). Generative ML and CSAM: Implications and Mitigations. Stanford Internet Observatory. https://purl.stanford.edu/jv206yg3793.

107 See Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. ArXiv. https://arxiv.org/pdf/2403.07918. (discussion of why open foundation models present an increase in marginal risk specifically for NCII).

108 Keller, M., & Dance, G. (2019, September 29). Last year, tech companies reported over 45 million online photos and videos of children being sexually abused—More than double what they found the previous year. NYTimes. https://www.nytimes.com/interactive/2019/09/28/us/child-sex-abuse.html.

109 Gendered Disinformation: Tactics, Themes, and Trends by Foreign Malign Actors - United States Department of State. (2023, April 12). United States Department of State. https://www.state.gov/gendered-disinformation-tactics-themes-and-trends-by-foreign-malign-actors/.

[110] EPIC comment attachment p. 3-4.

[111] Knibbs, K. (2024, January 26). Researchers Say the Deepfake Biden Robocall Was Likely Made With Tools From AI Startup ElevenLabs. https://www.wired.com/story/biden-robocall-deepfake-elevenlabs/.

[112] Elliott, V., & Kelly, M. (2024, January 23). The Biden Deepfake Robocall is Only the Beginning. https://www.wired.com/story/biden-robocall-deepfake-danger/.

[113] Suhasini, R. (2024, April 18). How A.I. Tools Could Change India’s Elections. https://www.nytimes.com/2024/04/18/world/asia/india-election-ai.html.

[114] Christopher, N. (2024, June 5). “The Near Future of Deepfakes Just Got Way Clearer.” The Atlantic. https://www.theatlantic.com/technology/archive/2024/06/india-election-deepfakes-generative-ai/678597/.

[115] Allyn, B. (2022, March 16). Deepfake video of Zelenskyy could be “tip of the iceberg” in info war, experts warn. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

[116] See, “Deepfake” of Biden’s Voice Called Early Example of US Election Disinformation. (2024, January 24). Voice of America. https://learningenglish.voanews.com/a/deepfake-of-biden-s-voice-called-early-example-of-us-election-disinformation/7455392.html; Hartmann, T. (2024, April 16). Viral deepfake videos of Le Pen family reminder that content moderation is still not up to par ahead of EU elections. www.euractiv.com. https://www.euractiv.com/section/artificial-intelligence/news/viral-deepfake-videos-of-le-pen-family-reminder-that-content-moderation-is-still-not-up-to-par-ahead-of-eu-elections/; Misinformation and disinformation. APA. (n.d.). https://www.apa.org/topics/journalism-facts/misinformation-disinformation. “False information deliberately intended to mislead.”

[117] Lohn, A. (2024, January 23). Deepfakes, Elections, and Shrinking the Liar’s Dividend. https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend.

[118] Josh A Goldstein, et al. (2024 February) How persuasive is AI-generated propaganda?. PNAS Nexus. https://doi.org/10.1093/pnasnexus/pgae034.

[119] J.A. Goldstein, et al. (2023) Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. ArXiv. https://arxiv.org/abs/2301.04246.

[120] See Access Now Comment at 2 “Nefarious actors can access them, remove built-in safety features, and potentially misuse them for malicious purposes, from malevolent actors creating disinformation to generate harmful imagery and deceptive, biased, and abusive language at scale.”

[121] J.A. Goldstein, et al. (2023) Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. ArXiv. https://arxiv.org/abs/2301.04246.

[122] Fergusson, G., & et al. (2023). Generative Harms Generative AI’s Impact & Paths Forward (p. 3). EPIC. https://epic.org/documents/generating-harms-generative-ais-impact-paths-forward/. Comment NTIA-2023-0009-0206.

[123] Kapoor, S., & Narayanan, A. (2023, June 16). How to Prepare for the Deluge of Generative AI on Social Media. https://knightcolumbia.org/content/how-to-prepare-for-the-deluge-of-generative-ai-on-social-media.

[124] Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. ArXiv. https://arxiv.org/pdf/2403.07918.

[125] Bateman, J., & Jackson, D. (2024). Countering Disinformation: Effectively An Evidence Based Policy Guide (p. 87). Carnegie Endowment. https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide.

[126] American Psychological Association. Misinformation and disinformation. https://www.apa.org/topics/journalism-facts/misinformation-disinformation.

[127] Perrigo, B. (2023, October 26). The Scientists Breaking AI to Make It Safer. Time. https://time.com/6328851/scientists-training-ai-safety/.

[128] Heikkilä, M. (2023, February 14). Why you shouldn’t trust AI search engines. Technology Review. https://www.technologyreview.com/2023/02/14/1068498/why-you-shouldnt-trust-ai-search-engines/.

[129] Bender, E., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? https://dl.acm.org/doi/pdf/10.1145/3442188.3445922.

[130] Chen, C. & Shu, K. (2023). Combating Misinformation in the Age of LLMs: Opportunities and Challenges. ArXiv preprint. https://doi.org/10.48550/arXiv.2311.05656.

[131] Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-127.

[132] Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-127.

[133] See generally: Florian Leiser, et al. (2023) From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users. In Proceedings of Mensch und Computer 2023 (MuC ‘23). Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/3603555.3603565; Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In Proceedings of the 5th International Conference on Conversational User Interfaces (CUI ‘23). Association for Computing Machinery, New York, NY, USA, Article 2, 1–10. https://doi.org/10.1145/3571884.3597144.

134 Neumeister, L. (2023, June 8). Lawyers blame ChatGPT for tricking them into citing bogus case law. https://apnews.com/article/artificial-intelligence-chatgpt-courts-e15023d7e6fdf4f099aa122437dbb59b.

135 Hsu, T. (2023, August 3). What Can You Do When A.I. Lies About You?

136 See generally, Wex Legal Dictionary and Encyclopedia. Discrimination. Legal Information Institute. (Last Updated November 2023) https://www.law.cornell.edu/wex/discrimination.

137 See American Civil Liberties Union, Center for American Progress, and the Leadership Conference on Civil and Human Rights Comment at 3 "These examples span the range of technology that may be deemed "artificial intelligence," and the emergence of applications such as chatbots and facial recognition technology underscores that both rudimentary and the most sophisticated AI technologies are already affecting civil rights, safety, and access to opportunities."

138 See generally: Xiang, C. (2023, March 22). The Amateurs Jailbreaking GPT Say They're Preventing a Closed-Source AI Dystopia. https://www.vice.com/en/article/5d9z55/jailbreak-gpt-openai-closed-source; Knight, W. (2023, December 5). A New Trick Uses AI to Jailbreak AI Models-Including GPT-4. https://www.wired.com/story/automated-ai-attack-gpt-4/; Ananya. (2024, March 19). AI image generators often give racist and sexist results: Can they be fixed? https://www.nature.com/articles/d41586-024-00674-9; Hofmann, V. (2024). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. https://arxiv.org/pdf/2403.00742.

139 Wiessner, D. (2024, February 21). Workday accused of facilitating widespread bias in novel AI lawsuit. https://www.reuters.com/legal/transactional/workday-accused-facilitating-widespread-bias-novel-ai-lawsuit-2024-02-21/.

140 Sadok, H., Sakka, F. & El Maknouzi, M. (2022). Artificial intelligence and bank credit analysis: A Review. Cogent Economics Finance, 10(1). https://doi.org/10.1080/23322039.2021.2023262.

141 Juhn, Y. et al (2022). Assessing socioeconomic bias in machine learning algorithms in health care : a case study of the HOUSES index. Journal of the American Medical Informatics Association, 29(7), 1142-1151. https://doi.org/10.1093/jamia/ocac052.

142 FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI. (2023, April 25). https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai.

143 See Connected Health Initiative Comment ("Successful creation and deployment of AI-enabled technologies which help care providers meet the needs of all patients will be an essential part of addressing this projected shortage of care workers. Policymakers and stakeholders will need to work together to create the appropriate balance between human care and decision-making and augmented capabilities from AI-enabled technologies and tools.").

144 See, Blumenfeld, J. (2023, December 22). NASA and IBM openly release Geospatial AI Foundation Model for NASA Earth observation data. Earthdata. https://www.earthdata.nasa.gov/news/impact-ibm-hls-foundation-model.

145 See Caleb Withers Comment at 9 ("Illustratively, the best coding models have either been, or been derived from, the most capable general-purpose foundation models, which are typically trained on curated datasets of coding data in addition to general training.").

146 National Institutes of Health (n.d.). Mission and Goals. Department of Health and Human Services. https://www.nih.gov/about-nih/what-we-do/mission-goals.

147 Bozeman, B. & Youtie, J. (2017). Socio-economic impacts and public value of government-funded research: Lessons from four UN National Science Foundation initiatives. Research Policy 46(8) 1387-1389. https://doi.org/10.1016/j.respol.2017.06.003.

148 See Center for Democracy & Technology, et al. (2024, March 25). RE: Openness and Transparency in AI Provide Significant Benefits for Society. https://cdt.org/wp-content/uploads/2024/03/Civil-Society-Letter-on-Openness-for-NTIA-Process-March-25-2024.pdf ("Open models promote economic growth by lowering the barrier for innovators, startups, and small businesses from more diverse communities to build and use AI. Open models also help accelerate scientific research because they can be less expensive, easier to fine-tune, and supportive of reproducible research.").

149 See ACT Association Comment ("For example, healthcare treatments and patient outcomes stand poised to improve disease prevention and conditions, as well as efficiently and effectively treat diseases through automated analysis of x-rays and other medical imaging.") and AI Policy and Governance Working Group at 5 ("Making foundation models more widely accessible, with appropriate safeguards, could drive innovation in research and business–capitalizing on the promise of public benefit. Study and use of state-of-the-art AI models, including Large Language Models and other models like AlphaFold, may lead to improvements in performance, safety, and scientific breakthroughs across various domains. These potential benefits can best be realized if other AI model assets, such as model training data, are also made widely available, and if models are not subject to restrictive licenses. Areas that stand to potentially gain from a commitment of ensuring the wide availability of AI tools and systems include, but are not limited to, innovation and novel applications in public health, biomedical research, and climate science that might be scaled in the public interest. Any decision to constrain the availability of dual-use open foundation models must carefully weigh and consider these potential societal and economic benefits.").

150 A generative AI tool to inspire creative workers. (2024, February 14). MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/a-generative-ai-tool-to-inspire-creative-workers.

151 Criddle, C. & Madhumita M. (2024, May 8). Artificial intelligence companies seek big profits from 'small' language models. Financial Times. https://www.ft.com/content/359a5a31-1ab9-41ea-83aa-5b27d9b24ef9.

[152] American Civil Liberties Union, Center for American Progress, and the Leadership Conference on Civil and Human Rights Comment at 4 "In addressing AI's risks for civil rights, safety, and access to opportunity, advocates, affected communities, and policymakers have championed a number of regulatory goals, including auditing and assessments, transparency, and explainability."; Hugging Face Comment at 6 "Maximally open systems, including training data, weights, and evaluation protocol, can aid in identifying flaws and biases. Insufficient documentation can reduce effectiveness"; Google Comment at 7 "Openly available models also enable important AI safety research and community innovation. A diverse pool of available models ensures that developers can continue to advance critical transparency and interpretability evaluations from which the developer community has already benefited. For example, researchers have demonstrated a method for reducing gender bias in BERT embeddings."

[153] Fair Housing Testing Program. (2015, August 6). Justice.gov. https://www.justice.gov/crt/fair-housing-testing-program-1 ("In 1991, the Civil Rights Division established the Fair Housing Testing Program within the Housing and Civil Enforcement Section, which commenced testing in 1992.").

[154] Other sections in this Report also contain references to these topics.

[155] OpenAI's GPT-4, for example, cost around $100 million USD to train. See Knight, W. (2023, April 17). OpenAI's CEO Says the Age of Giant AI Models Is Already Over. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

[156] Birchler, U., & Bütler, M. (2007). Information Economics (Routledge Advanced Texts in Economics and Finance) (1st Edition).

[157] Jones, R., & Mendelson, H. (2011). Information Goods vs. Industrial Goods: Cost Structure and Competition. Management Science, 57(1), 164–176. http://www.jstor.org/stable/41060707.

[158] Nix, N., & et al. (2024, March 10). Silicon Valley is pricing academics out of AI research. https://www.washingtonpost.com/technology/2024/03/10/big-tech-companies-ai-research/.

[159] Lee, K., & et al. (2024, March 12). Building Meta's GenAI Infrastructure. https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/. See also, Clark, J. (2024, March 25). Import AI 366: 500bn text tokens; Facebook vs Princeton; why small government types hate the Biden EO. https://jack-clark.net/2024/03/25/import-ai-366-500bn-text-tokens-facebook-vs-princeton-why-small-government-types-hate-the-biden-eo/.

[160] Hays, K. (2024, January 19). Zuck's GPU Flex Will Cost Meta as Much as 18 Billion by the end of 2024. https://www.businessinsider.com/mark-zuckerberg-ai-flex-meta-nvidia-gpus-2024-1.

[161] Meta aims to have 350,000 NVIDIA H100 GPUs by the end of the year. If each one costs $20,000 (a modest estimate according to the Business Insider article above), the total cost will be $7B. This does not include Meta's other computing resources, or the money required for datasets, human resources, or other requirements like energy.

[162] See, e.g., Vipra, J., & Korinek, A. (2023). Market concentration implications of foundation models: The Invisible Hand of ChatGPT. Brookings at 9-24 (analyzing the economics of foundation models). See also CDT Comment at 5. Cf. Economic Report of the President. p.280 (2024). The White House. https://www.whitehouse.gov/wp-content/uploads/2024/03/ERP-2024.pdf. ("In other cases, however, some combination of high entry costs, data availability, and network effects may drive markets toward having only a small number of players. Markets for generative AI products, which require huge amounts of data and computing power to train, may be particularly prone to this issue, with some even suggesting that such markets may naturally trend toward monopoly[. . .]." (internal citation omitted).

[163] See, Kapoor, S.et al., (2024). On the Societal Impact of Open Foundation Models. at 5. ArXiv. https://arxiv.org/pdf/2403.07918.

[164] See generally Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations. Hugging Face. https://arxiv.org/pdf/2302.04844. See also Hugging Face Comment at 10-15.

[165] See, e.g., Alliance for Trust in AI Comment at 5 ("While available model weights may make it easier to develop advanced AI, there are still significant barriers to run and modify large or advanced models. It is not clear whether the model weights themselves provide enough information to end users to significantly change what they can do or develop themselves."); Intel Comment at 8 ("[A]lmost all innovation in AI to-date has been due to openly available infrastructure [. . .]" (beyond just model weights to include "architecture and dataset transparency.")). See also RAND Comment at 2 ("Whether [access to open foundation models] will be enough to maintain a competitive market for foundation model based products or services in general will depend on the price to develop and the performance of open models compared with closed models and on how the economics of fine-tuning, adapting, and serving foundation models differs in a particular business application between large and small companies.") (internal citation omitted).

[166] See, e.g., Engine Comment at 3 ("Moreover, the extent of openness matters. Whether open source AI resources, for example, include detailed documentation, have publicly available model weights, or license-based restrictions can impact how useful those resources are for startups. Policymakers should be very clear-eyed about consequences for startups and innovation of adding policy-related barriers to these resources."); Public Knowledge Comment at 11 ("Open source model weights, commercially available data warehouses, and public compute resources would enable many new model developers to use the data to develop and train new models. In addition, foundation models and APIs could also be opened, so that developers have reliable access to these resources.") (internal citation omitted); ACLU et al. Comment at 9 ("The potential promise of 'open' AI is that it may allow increased competition and customization of AI models, disrupting the potential concentration developing in the advanced AI market. However, this competition will only exist if 'open' AI models are able to be hosted at the scale necessary for success.") (quotation marks in original). Cf. IBM Comment at 5 ("The most obvious benefit of an open ecosystem is that it lowers the barrier to entry for competition and innovation. By making many of the technical resources necessary to develop and deploy AI more readily available, including model weights, open ecosystems enable small and large firms alike, as well as

research institutions, to develop new and competitive products and services without steep, and potentially prohibitive, upfront costs."). See also Widder, D., & et al. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. at 7. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807. ("Access to compute presents a significant barrier to reusability for even the most maximally 'open' AI systems, because of the high cost involved in both training and running inferences on large-scale AI models at scale (i.e. instrumenting them in a product or API for widespread public use).") (quotation marks in original); Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource. at v. (2023). https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf. ("The [National AI Research Resource] should comprise a federated set of computational, data, testbed, and software resources from a variety of providers, along with technical support and training, to meet the needs of [its] target user base."). Cf. Economic Report of the President. at 281. (2024). The White House. https://www.whitehouse.gov/wp-content/uploads/2024/03/ERP-2024.pdf. ("Similarly, freely available and portable data may encourage a competitive landscape and ensure that gains from data are widely distributed.").

167 See, e.g., ACLU et al. Comment at 9 ("Currently, the major commercial cloud computing vendors allow other AI models, including 'open' AI models, to be hosted on their cloud computing services. But there is no requirement for any major commercial cloud computing vendors to allow 'open' AI models to be hosted on their services, and the potential for self-preferencing may make the use of non-native AI models more difficult or expensive.") (internal citation omitted) (quotation marks in original). Open models also benefit the cloud computing market, dominated by Amazon, Google, and Microsoft, which also shows anticompetitive and cumulative advantage features. See generally, e.g., Narechania, T., & Sitaraman, G. (2023). Working Paper Number 24-8. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4597080. See also Paul, K. (2023, July 18). Meta opens AI model to commercial use, throwing nascent market into flux. https://www.reuters.com/technology/meta-opens-ai-model-commercial-use-throwing-nascent-market-into-flux-2023-07-18/. ("Asked why Microsoft would support an offering that might degrade OpenAI's value, a Microsoft spokesperson said giving developers choice in the types of models they use would help extend its position as the go-to cloud platform for AI work.").

168 The Open Source Definition. (2024, February 16). https://opensource.org/osd.

169 Hoffmann, M., & et al. (2024). The Value of Open Source Software (Harvard Business School Strategy Unit Working Paper No. 24-038). Harvard Business School. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4693148.

170 See, e.g., CDT Comment at 2-4.

171 Blind, K., & Schubert, T. (2023). Estimating the GDP effect of Open Source Software and its complementarities with R&D and patents: Evidence and policy implications. The Journal of Technology Transfer, 49:466–491. https://link.springer.com/article/10.1007/s10961-023-09993-x.

172 West, J., & Gallagher, S. (2006). Challenges of open innovation: The paradox of firm investment in open-source software. 36(3), 319–331.

173 See, e.g., Mozilla Comment at 12 ("As Widder, West, and Whittaker have argued, promoting openness in AI alone is not sufficient for creating a more competitive ecosystem. There are also risks of openness being co-opted by big industry players, and a long track record of companies drawing significant benefits from open source technology without re-investing into the communities that have developed those technologies."), referencing Widder, D., & et al. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. at 6. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807. See also ACLU et al. Comment at 6 ("Further compounding the complexity around 'open' AI is the fact that it is not always easy to separate 'openness' from the business interests of large AI developers, who may benefit from open innovation on their platforms and may later withdraw commitments to openness after the benefits have reached a critical mass, knowing that smaller developers are unlikely to have the resources necessary to independently compete.") (quotation marks in original) (internal citations omitted).

174 Paul, K. (2023, July 18). Meta opens AI model to commercial use, throwing nascent market into flux. https://www.reuters.com/technology/meta-opens-ai-model-commercial-use-throwing-nascent-market-into-flux-2023-07-18/.

175 Yao, D. (2023, July 27). Meta to Charge for Llama 2 After All – If You're a Hyperscaler. https://aibusiness.com/nlp/meta-to-charge-for-llama-2-after-all-if-you-re-a-hyperscaler.

176 Widder, D., & et al. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. at 13. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807.

177 See Engler, A. (2021, August 10). How open-source software shapes AI policy. https://www.brookings.edu/articles/how-open-source-software-shapes-ai-policy/. ("In fact, for Google and Facebook, the open sourcing of their deep learning tools (Tensorflow and PyTorch, respectively), may have the exact opposite effect, further entrenching them in their already fortified positions. While [open source software] is often associated with community involvement and more distributed influence, Google and Facebook appear to be holding on tightly to their software. [. . .] [T] these companies are gaining influence over the AI market through OSS, while the OSS AI tools not backed by companies, such as Caffe and Theano, seem to be losing significance in both AI research and industry. By making their tools the most common in industry and academia, Google and Facebook benefit from the public research conducted with those tools, and, further, they manifest a pipeline of data scientists and machine learning engineers trained in their systems.") (internal hyperlink omitted).

178 Staff in the Federal Trade Commission Office of Technology. (2024, July 10). On Open-Weights Foundation Models. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/open-weights-foundation-models.

179 Scharre, P. (2024, March 13). Future-Proofing Frontier AI Regulation. https://www.cnas.org/publications/reports/future-proofing-frontier-ai-regulation.

180 See, Melton, M. (2024, February). Generative AI startup Latimer, known as the "BlackGPT", will launch a new bias detection tool and API. Business Insider. https://www.businessinsider.com/latimer-ai-api-launch-bias-detection-tools-llm-2024-2.

181 See, Rishi Bommasani et al. Comment at 5 ("Open foundation models promote competition in some layers of the AI stack. Given the significant capital costs of developing foundation models, broad access to model weights and greater customizability can also reduce market concentration by enabling greater competition in downstream markets. However, open foundation models are unlikely to reduce market concentration in the highly concentrated upstream markets of computing and specialized hardware providers.") (internal citation omitted); RAND Comment at 2 ("Whether [access to open foundation models] will be enough to maintain a competitive market for foundation model based products or services in general will depend on the price to develop and the performance of open models compared with closed models and on how the economics of fine-tuning, adapting, and serving foundation models differs in a particular business application between large and small companies.") (internal citation omitted). See also, Hugging Face Comment at 9 ("Additionally, open-weight models have been customized and adapted to run on a greater variety of infrastructure, including individual GPUs and even CPUs, reducing points of market concentration with cloud providers and reducing costs for procurement.") (referencing Ggerganov / llama.cpp. (n.d.). https://github.com/ggerganov/llama.cpp and Hood, S. (2023, December 14). llamafile: Bringing LLMs to the people, and to your own computer. https://future.mozilla.org/news/introducing-llamafile/); CDT Comment at 9 ("Importantly, the innovation in developing smaller and more powerful models, often based directly on much larger models, is not just important in terms of competition and innovation. It is also important because some models such as Mistral 7B are now small enough to run locally on an end-user's laptop or even a phone, mitigating the need for a cloud-based provider at all.") (internal citation omitted).

182 See, e.g., RAND Comment at at 2 ("Open foundation models may reduce market concentration. When smaller actors can access open foundation models, they can avoid the large expense of developing their own models and can therefore compete with large tech companies in adapting the foundation model to a particular business context.") (internal citation omitted), Engine Comment at 3 ("Openness in AI helps alleviate costs associated with the expensive parts of building models, leaving startups to focus their limited resources on their core and differentiating innovation."); The Abundance Institute Comment at 4 ("In particular, the high costs of compiling data and purchasing compute to train foundational models are a significant barrier to model training. Sharing model weights eliminates this cost barrier, broadening access and enabling users that would otherwise simply be priced out of building their own AI stack.").

183 See, e.g., a16z Comment at 11 ("Open Models increase competition in the development and improvement of foundation models because they do not restrict the use of AI to gatekeeper companies with the most market power or resources. This accessibility increases the prospect of competition and allows for participation by developers who may otherwise have been boxed out of working with AI due to their lacking the requisite access or resources that are necessary components to working within a closed ecosystem."); CSET Comment at 4 ("Downloadable weights [. . .] may reduce the concentration of power in the AI industry because the original developers do not control access to the models."); Intel Comment at 9 ("Open model weights also spur more startups and innovations, enabling startups to quickly prototype without access to immense capital, fostering a more competitive landscape."). Cf. Public Knowledge Comment at 10 ("Dominant companies often utilize gatekeeper power to further their own market power and cut off new entrants from the chance to compete. Open technologies may serve to counteract this exclusionary conduct and lower barriers to entry for innovative, up-and-coming rivals. Historically, we've already seen how open access to technology patents had competitive benefits, leading to a wellspring of innovative products.").

184 See, e.g., Mozilla Comment at 12 ("The increased availability of 'open' alternatives in the AI market can support competition by reducing switching costs as relying on specific proprietary model APIs or platform ecosystems (like those offered by the leading cloud service providers) can create lock-in effects for customers, both in the private and public sector.") (quotation marks in original). Open models allow companies to switch seamlessly between baseline models without added costs. Proprietary models introduce the threat of "lock-in effects," where a company has built a product around a certain API or provider and then cannot transfer to a new model without rebuilding their entire product. Many technology corporations have established verticals with certain cloud service providers and data collection infrastructures, and a company cannot easily exit this vertical. Conversely, when companies build on open models, they have access to those model weights forever and can switch between cloud providers and other vendors with ease.

185 See Kapoor, S. (2024). On the Societal Impact of Open Foundation Models. at 5. ArXiv. https://arxiv.org/pdf/2403.07918.

186 United Nations General Assembly, Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development. March 11, 2024. A/78/L.49. https://www.undocs.org/A/78/L.49.

187 See ACLU et al. Comment at 10 ("As seen in other technological contexts, diffusing market concentration, especially over gateway or bottleneck facilities, can increase the diversity of voices, including for marginalized communities.") (internal citation omitted); GitHub Comment at 10 ("An expanded developer base, particularly outside of a small set of companies located in a few major tech hubs, supports diversity of identity and perspective in the ecosystem.").

188 See CDT Comment at 7 ("[Open Foundation Models] are already driving innovation across the ecosystem as tens or hundreds of thousands of businesses begin adapting model capabilities to their own use cases and customer needs in a wide variety of contexts."); Phase 2 at 3 ("Making foundation model weights widely available lowers barriers to entry and enables a broader range of companies to develop AI applications. This is particularly beneficial for startups and small businesses that lack the resources to develop foundation models from scratch. Open models level the playing field and ensure the economic gains from AI are widely distributed. We expect this to drive competition and innovation in sectors like healthcare, education, and marketing as more players are able to leverage AI to build groundbreaking products and services.").

189 Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. ArXiv. https://arxiv.org/pdf/2403.07918.

[190] See, e.g., CTA Comment at 5 "Because [open weight models] have lower barriers to entry (e.g., cost, expertise), they are more accessible to the general public. Leveraging input and feedback from the broader AI community of researchers and users can help identify and mitigate bugs, biases, and safety issues that may otherwise go unnoticed, ultimately leading to better performing and safer AI products. This lower barrier to entry can help to drive AI research and development by academics or other subject matter experts, enabling communities with bespoke datasets and unique needs to form around specific platforms or industry sectors." (citing Elizabeth Seger, et al. (Sept. 29, 2023). Open-Sourcing Highly Capable Foundation Models. https://arxiv.org/pdf/2311.09227.pdf; Ly, J. (2024, March 12). Open Foundation Models: Implications of Contemporary Artificial Intelligence. Center for Security and Emerging Technology. https://cset.georgetown.edu/article/open-foundation-models-implications-of-contemporary-artificial-intelligence/.

[191] There may also be ways to achieve similar benefits to research and development through methods other than making model weights widely available. See, e.g., RAND Comment at 4 ("Structured access is an alternative approach that can provide users with many of the benefits of making foundation model weights widely accessible while reducing some of the risks."); Anthony Barrett Comment at 15 ("[B]enefits of broad independent evaluation for improving the safety, security, and trustworthiness of AI are not necessarily best supported by making model weights widely available. Those benefits can also be achieved by facilitating safe and protected independent researcher access.").

[192] See, e.g., Hugging Face Comment at 9 ("Robust innovation on both performance and safety questions requires scientific rigor and scrutiny, which is enabled by openness and external reproducibility. Supporting that research requires sharing models to validate findings and lower the barrier to entry for participation given the growing resource gap between researchers in different institutions.") (internal citations omitted); Center for Democracy & Technology, & et al. (March, 25, 2024). RE: Openness and Transparency in AI Provide Significant Benefits for Society at 2. https://cdt.org/wp-content/uploads/2024/03/Civil-Society-Letter-on-Openness-for-NTIA-Process-March-25-2024.pdf. ("Open models also help accelerate scientific research because they can be less expensive, easier to fine-tune, and supportive of reproducible research."). See Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. at 19. ArXiv. https://arxiv.org/pdf/2403.07918 (listing examples of research done using open foundation models).

[193] See, e.g., RAND Comment at 3 ("Making foundation model weights accessible helps uncover vulnerabilities, biases, and potentially dangerous capabilities. With a wider set of eyes examining these models, there is a higher likelihood of identifying and addressing issues that might have been overlooked by the original developers, as is the case with open-source software broadly. This scrutiny is useful for developing AI systems that are secure, fair, and aligned with societal values. The detection and mitigation of biases in AI models, for instance, are critical steps toward ensuring that AI technologies do not perpetuate or exacerbate social inequalities."); IBM Comment at 6 ('In some contexts, AI safety can also depend on the ability for diverse stakeholders to scrutinize and evaluate models to identify any vulnerabilities, identify undesirable behaviors, and ensure they are functioning properly. However, without 'deep access,' which includes access to model weights, these evaluations will be severely limited in their effectiveness.") (internal citation omitted); CDT Comment at 10-14 (explaining the "black-box" methods of auditing for closed foundation models versus "white-box" methods of auditing for open foundation models). Cf. Engler, A. (2021, August 10). How open-source software shapes AI policy. https://www.brookings.edu/articles/how-open-source-software-shapes-ai-policy/ ("Similarly, open-source AI tools can enable the broader and better use of ethical AI. Open-source tools like OSS like IBM's AI Fairness 360, Microsoft's Fairlearn, and the University of Chicago's Aequitas ease technical barriers to detecting and mitigating AI bias. There are also open-source tools for interpretable and explainable AI, such as IBM's AI Explainability 360 or Chris Molnar's interpretable machine learning tool and book, which make it easier for data scientists to interrogate the inner workings of their models.").

[194] See GitHub Comment at 9 ("To-date, researchers have credited [open source and widely available AI] models with supporting work to [. . .] advance the efficiency of AI models enabling them to use less resources and run on more accessible hardware.") (citing Tim Dettmers, et al., "QLoRA: Efficient Finetuning of Quantized LLMs," ArXiv, May 23, 2023, https://arxiv.org/abs/2305.14314 and its associated repository at Artidoro / qlora. (n.d.). https://github.com/artidoro/qlora).. https://doi.org/10.48550/arXiv.2202.07646).

[195] See, e.g., Intel Comment at 8-9 ("Open model weights are likely going to aid researchers to find impactful and beneficial use cases of AI that will be overlooked by narrow and immediate commercial interests of proprietary model vendors. An example of this is applying leading-edge AI principles to open scientific problems."); OTI Comment at 16 ("One of the key benefits of a healthy ecosystem characterized by a prevalence of open models is that many people can learn how the technology works. This enables technologists and community leaders to partner in ways that are tailored to address specific community needs and implement community-driven solutions. Relatedly, open-source projects can also be used to fill technological gaps that aren't being met in the private sector.") (internal citations omitted).

[196] See MLCommons Comment at 3 ("Models with open weights have played a central role in developing widely trusted benchmarks that have been used to evaluate and measure AI models, and in doing so have helped drive progress in AI. GLUE, BigBench, Harness, HELM and openCLIP Benchmark are all examples of widely used benchmarks that have helped researchers and developers measure progress in the development of AI models.") (internal citations omitted).

[197] See EleutherAI Comment at 24 ("Even for researchers in industrial labs such as Google, open models can enable research on model safety that would not otherwise be possible: in an earlier revision of Quantifying Memorization Across Neural Language Models, Carlini et al. state that their research on harmful memorization in language models "would not have been possible without EleutherAI's complete public release of The Pile dataset and their GPT-Neo family of models.") (internal citations omitted). See also Zou, A. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. https://arxiv.org/abs/2307.15043.

[198] See Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. at 4. ArXiv. https://arxiv.org/pdf/2403.07918. (referencing research that

requires no safety filters). See also CDT Comment at 9 ("Furthermore, [open foundation models] enable a variety of AI research not enabled by closed foundation models, including research around AI interpretability methods, security, model training and inference efficiency, and the public development of robust watermarking techniques.") (listing examples) (internal citations omitted). See also Kapoor, S., & Narayanan, A. (2023, March 22). OpenAI's policies hinder reproducible research on language models. https://www.aisnakeoil.com/p/openais-policies-hinder-reproducible. To be clear, the benefits in this space are not a zero-sum game. One may need access to both open weight models and "closed" foundation models. See, e.g. MLCommons Comment at 3 (describing the limitations of relying solely on models with open weights or models with closed weights to evaluate models and urging simultaneous use).

199 See, e.g., CSET Comment at 11 ("Most current [Biological Design Tools] are open models developed by academic labs. The life sciences community places a high value on scientific transparency and openness, and tends to favor open sharing of resources [. . .] Shifting away from the open sharing of model weights would also require additional resources, as many academic researchers do not have the time, funding, and infrastructure to set up and maintain an API.") (internal hyperlink omitted).

200 See, e.g., Miller, K. (2024, March 12). Open Foundation Models: Implications of Contemporary Artificial Intelligence. https://cset.georgetown.edu/article/open-foundation-models-implications-of-contemporary-artificial-intelligence/. ("Actors may opt to use open models instead of paying for access to closed models, which may reduce the revenue of developers and disincentivize investments in capital-intensive R&D.").

201 See, e.g., CSET Comment at 10 ("More research is needed to determine what types of research are enabled by open weights, and how that may allow more entrants into the market. Many prospective entrants may lack resources, and it is unclear the extent to which resource constraints may limit the benefits of open models to R&D. Actors may lack the data to fine-tune open models, or lack the compute to use or experiment with open models rigorously and at scale (although resources provided through the NAIRR pilot may help alleviate resource constraints).") (internal hyperlinks omitted).

202 See Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. at 4. ArXiv. https://arxiv.org/pdf/2403.07918.

203 See Kapoor, S. et al., (2024). On the Societal Impact of Open Foundation Models. at 4. ArXiv. https://arxiv.org/pdf/2403.07918. These researchers also note that "new research directions such as merging models might allow open foundation model developers to reap some of these benefits (akin to open source software)." (internal citation omitted).

204 See Kleinberg, J., & Raghavan, M. (2020). Algorithmic monoculture and social welfare. Proceedings of the National Academy of Sciences of the United States of America, 118(22) at 1. https://www.pnas.org/doi/epdf/10.1073/pnas.2018340118.

205 See generally id. Kleinberg and Raghavan highlight several concerns with algorithmic monoculture: 1) risk of severe harm in monoculture systems due to unexpected shocks, and 2) decrease in decision-making quality across the board. See also, e.g., a16z Comment at 20 ("Algorithmic monocultures resulting from reliance on a few Closed Models can create resilience problems and generate systemic risk. If those models are compromised, the impacts could be widespread and pervasive."). Cf. Vipra, J., & Korinek, A. (2023). Market concentration implications of foundation models: The Invisible Hand of ChatGPT. at 25. Brookings. ("Foundation models will likely be integrated into production and delivery processes for goods and services across many sectors of the economy. We can imagine one foundation model in its fine-tuned versions powering decision-making processes in search, market research, customer service, advertising, design, manufacturing, and many more. If foundation models are integrated into a growing number of economic activities, then widespread, cross-industrial applications mean that any errors, vulnerabilities, or failures in a foundation model can threaten a significant amount of economic activity, producing the risk of systemic economic effects.").

206 See, e.g., CDT Comment at 5 ("[W]hen many different decisionmakers and service providers rely on the same systems, there can be a trend toward 'algorithmic monoculture' whereby systemic exclusion of individuals or groups in AI-driven decisionmaking occurs across the ecosystem") (citing Rishi Bommasani et al., "Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?," ArXiv, November 25, 2022, https://arxiv.org/abs/2211.13972. [perma.cc/F7JB-3AK3]).

207 See, e.g., Mozilla Comment at 12 ("Additionally, concentrating cutting-edge research in ever-fewer research labs may also exacerbate phenomena like algorithmic monoculture and entrench (or increase the 'stickiness' of) existing technological paradigms at the expense of pursuing new research directions) (quotation marks in original), citing Fishman, N., & Hancox-Li, L. (2022). Should attention be all we need? The epistemic and ethical implications of unification in machine learning. https://dl.acm.org/doi/abs/10.1145/3531146.3533206. and Hooker, S. (2020). The Hardware Lottery. ArXiv. https://arxiv.org/abs/2009.06489. See also Fishman, N., & Hancox-Li, L. (2022). Should attention be all we need? The epistemic and ethical implications of unification in machine learning. at 14. https://dl.acm.org/doi/abs/10.1145/3531146.3533206.

208 Faverio, M. and Tyson, A. (2023). What the data says about Americans' views of artificial intelligence. Pew Research Center. https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/.

209 Sartori, L. and Theodorou, A (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics and Information Technology, Vol 24, 4. https://link.springer.com/10.1007/s10676-022-09624-3.

210 See, e.g., Uber Comment at 2 ("…open-source models create a more level playing field and lower barriers to entry for AI use, ensuring that more individuals and organizations can access and improve upon existing technology.").

211 Verma, P. & Zakrzewski, C. (April 23, 2023). AI deepfakes threaten to upend global elections. No one can stop them. Washington Post. https://www.washingtonpost.com/technology/2024/04/23/ai-deepfake-election-2024-us-india/.

212 Singer, N. (April 8, 2023). Teen Girls Confront an Epidemic of Deepfake Nudes in Schools. New York Times. https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html.

213 Sadok, H., Sakka, F. & El Maknouzi, M. (2022). Artificial intelligence and bank credit analysis: A Review. Cogent Economics Finance, 10(1). https://doi.org/10.1080/23322039.2021.2023262.

214 Juhn, Y. et al (2022). Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. Journal of the American Medical Informatics Association, 29(7), 1142-1151. https://doi.org/10.1093/jamia/ocac052.

215 Polonski, V. (2018). AI is convicting criminals and determining jail time, but is it fair? World Economic Forum: Emerging Technologies. https://www.weforum.org/agenda/2018/11/algorithms-court-criminals-jail-time-fair/.

216 See, e.g., Hugging Face Comment at 10 ("In most cases, the risks associated with open-weight models are broadly similar to any other part of a software system (with or without AI components), and are similarly context-dependent.") (citations and emphasis omitted); OpenAI Comment at 3 ("As AI models become even more powerful and the benefits and risks of their deployment or release become greater, it is also important that we be increasingly sophisticated in deciding whether and how to deploy a model. This is particularly true if AI capabilities come to have significant implications for public safety or national security. The future presence of such 'catastrophic' risks from more advanced AI systems is inherently uncertain, and there is scholarly disagreement on how likely and how soon such risks will arise.") (quotation marks in original); Mozilla Comment at 11 ("There is so much unknown about the benefits of AI, and policymakers must not ignore this."); Microsoft Comment at 15 ("Moreover, even when model and application developers take all reasonable precautions to assess and mitigate risks, mitigations will fail, unmitigated risks will be realized, and unknown risks will emerge. These risks could range from generating harmful content in response to a malicious prompt to the intentional exfiltration of an advanced AI model by a nation state actor."). Cf. JHU Comment at 9 ("Given the uncertain nature of current and future open model capabilities, and the importance of open software, we are not suggesting that the DOC should impose export controls on dual-use foundation models today. Rather, the risks posed by open, biologically capable dual-use foundation models are grave enough for the US government to prepare such policy options so they can be deployed when and if they become relevant.").

217 Web's inventor gets a knighthood. (December 31, 2003). BBC News. http://news.bbc.co.uk/2/hi/technology/3357073.stm.

218 Cai, Z. et al. (2023). Associations Between Problematic Internet Use and Mental Health Outcomes of Students : A Meta-analytic Review. Quantitative Review 8(45). https://doi.org/10.1007/s40894-022-00201-9; Katz, L. (June 18, 2020). How tech and social media are making us feel lonelier than ever. CNET. https://www.cnet.com/culture/features/how-tech-and-social-media-are-making-us-feel-lonelier-than-ever/.

219 Nestor Maslej, et al. Artificial Intelligence Index Report 2024. (2024 April). https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf at 14 ("Between 2010 and 2022, the total number of AI publications nearly tripled, rising from approximately 88,000 in 2010 to more than 240,000 in 2022.").

220 Nestor Maslej, et al. Artificial Intelligence Index Report 2024. (2024 April). https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf at 46 ("Until 2014, academia led in the release of machine learning models. Since then, industry has taken the lead.").

221 Marchant, G. (2020). Governance of Emerging Technologies as a Wicked Problem. Vanderbilt Law Review 73(6). https://scholarship.law.vanderbilt.edu/vlr/vol73/iss6/8/; Holtel, S. (2016). Artificial Intelligence Creates a Wicked Problem for the Enterprise. Procedia Computer Science 99, 171-180. https://doi.org/10.1016/j.procs.2016.09.109; Also see generally, Head, B. & Alford, J. Wicked Problems: Implications for Public Policy and Management. Administration & Society 47(6). https://doi.org/10.1177/0095399713481601; Walker, W., Marchau, V. & Swanson, D. (2010). Addressing deep uncertainty using adaptive policies: Introduction to section 2. Technological Forecasting and Social Change 77(6), 917-923. https://doi.org/10.1016/j.techfore.2010.04.004.

222 The National Artificial Intelligence Advisory Committee (NAIAC) (2023). RECOMMENDATIONS: Generative AI Away from the Frontier. https://ai.gov/wp-content/uploads/2023/11/Recommendations_Generative-AI-Away-from-the-Frontier.pdf; Kwakkel, J., Walker, W.,& Haasnoot, M. (2016). Coping with the Wickedness of Public Policy Problems: Approaches for Decision Making under Deep Uncertainty. Journal of Water Resources Planning and Management 142(3). https://doi.org/10.1061/(ASCE)WR.1943-5452.0000626.

223 Holtel, S. (2016). Artificial Intelligence Creates a Wicked Problem for the Enterprise. Procedia Computer Science 99, 171-180. https://doi.org/10.1016/j.procs.2016.09.109; Berente, N., Kormylo, C., & Rosenkranz, C. (2024). Test-Driven Ethics for Machine Learning. Communications of the ACM 67(5), 45-49. https://cacm.acm.org/opinion/test-driven-ethics-for-machine-learning/.

224 Liu, H. & Maas, M. (2021). 'Solving for X?' Towards a Problem-Finding Framework to Ground Long-Term Governance Strategies for Artificial Intelligence. Futures 126(22). https://doi.org/10.1016/j.futures.2020.102672.

225 Maier, H.R. et al (2016). An uncertain future, deep uncertainty, scenarios, robustness and adaptation : How do they fit together? Environmental Modelling & Software 81, 154-164. https://doi.org/10.1016/j.envsoft.2016.03.014; RAND. Robust Decision Making. Water Planning for the Uncertain Future. https://www.rand.org/pubs/tools/TL320/tool/robust-decision-making.html.

226 Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. Global Environmental Change, 23(2), 485–498. https://doi.org/10.1016/j.gloenvcha.2012.12.006.

227 Rand Comment at 4 (“The most common approach to structured access is to create flexible application programming interfaces (APIs) that allow researchers, small businesses, or the public to access the model.”).

228 Anthony Barrett Comment at 3 (“Foundation model developers that plan to provide downloadable, fully open, or open source access to their models should first use a staged-release approach (e.g., not releasing parameter weights until after an initial secured or structured access release where no substantial risks or harms have emerged over a sufficient time period), and should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk management has been established, including for safety risks and risks of misuse and abuse.”) (internal citation omitted).

229 Johns Hopkins Center for Health Security at 6 (“…none of the small studies in the field so far have evaluated how much dual-use foundation models purposefully trained on relevant data (eg, virology literature) will marginally improve bioweapons development or assessed the interaction between LLMs and BDTs. 24 Nor, to our knowledge, have there been any published evaluations of the marginal benefit BDTs like Evo or RFdiffusion could play in bioweapons design.”).

230 Mozur, P. et al. (2024, February 21). China’s Rush to Dominate A.I. Comes With a Twist: It Depends on U.S. Technology. NYTimes. https://www.nytimes.com/2024/02/21/technology/china-united-states-artificial-intelligence.html.

231 National Association of Manufacturers Comment at 2 (“The availability of model weights allows independent examination of a model to ensure it is fit for purpose and to identify and mitigate its vulnerabilities.”). At the same time, the benefit of transparency may be relative to the availability of other components and resources. See, e.g., CSET Comment at 16 (“Models with publicly available weights fit along a spectrum of openness, and where they fit depends on the accessibility of their components [. . .] More research is needed to gauge how different degrees of access and transparency can impact the ability to scrutinize or evaluate open models. For example, many open model come with documentation and model cards, but the level of detail in these documents can vary dramatically, and they can enable (or not enable) different degrees of evaluation.”); Databricks Comment at 10 (“Making the model code widely available in addition to the model weights provides the benefits of incremental transparency in evaluation the model[. . .].”).

232 See, e.g., EleutherAI Institute Comment at 24 (“Open-weights models allow more researchers than just the small number of at industry labs to investigate how to improve model safety, improving the breadth and depth of methods that can be explored, and also allows for a wider demographic of researchers or auditors of safety.”); Databricks Comment at 5 (“The biggest risks Databricks sees are the risks that would be created by prohibiting the wide availability of model weights: i.e., the risks to economic productivity benefitting a larger swath of society, innovation, science, competition, and AI transparency if Open DUFMs were not widely available.”).

233 See, e.g., EleutherAI Institute Comment at 24 (listing examples of research facilitated by “open-weights foundation models.”); Rishi Bommasani et al. Comment at 3 (“Model weights are essential for several forms of scientific research across AI interpretability, security, and safety) ; CDT Comment at 8 (“Researchers used the model weights of Mistral 7B [. . .] to decrease the computational power required for fine-tuning the model for downstream tasks by a factor of ten.”).

234 For example, Stable Diffusion 3. (2023, February 22). https://stability.ai/news/stable-diffusion-3. “suite of models currently ranges from 800M to 8B parameters.”

235 See, e.g., CDT Comment at 33-40; The Abundance Institute Comment at 7 (“Like object code, model weights communicate information to a computer – in this case, a computer running an inference engine. [. . .] People and organizations who wish to publish such model weights have a protected speech interest in doing so.”); Mozilla Comment at 10 n.2 (“Further, as U.S. courts have held multiple times, computer source code must be viewed as expressive for First Amendment purposes. [. . .] A similar argument could be made about the importance of protecting the sharing of information about model weights and other AI components.”). Cf. G.S. Hans Comment at 2 (“Regulated AI companies may rely upon the reasoning of [Bernstein v. U.S.] to argue that export restrictions on model weights violate the First Amendment.”). But see Rozenshtein, A. (April 4, 2024). There Is No General First Amendment Right to Distribute Machine-Learning Model Weights. Lawfare. There may also be other constitutional challenges regarding openness in AI models beyond restriction of model weights. See generally G.S. Hans Comment (outlining a range of potential First Amendment challenges related to government requirements on transparency, content moderation, and other topics.).

236 Engler, A. (Jan. 22, 2024). The case for AI transparency requirements. https://www.brookings.edu/articles/the-case-for-ai-transparency-requirements/.

237 Greene, T., & et al. (2022). Barriers to academic data science research in the new realm of algorithmic behaviour modification by digital platforms. Nature Machine Intelligence, 4, 323–330.

238 Gorwa, R., & Veale, M. (2023, November 21). Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries. ArXiv.org. https://export.arxiv.org/abs/2311.12573v1.

239 National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). (2023). https://doi.org/10.6028/NIST.AI.100-1.

240 Longpre, S., & et al. (2024). A Safe Harbor for AI Evaluation and Red Teaming. ArXiv. https://arxiv.org/pdf/2403.04893.

241 Google Comment at 3.

242 National Telecommunications and Information Administration (2024). Artificial Intelligence Accountability Policy Report. https://www.ntia.gov/sites/

default/files/publications/ntia-ai-report-final.pdf.

243 IBM Comment at 4.

244 AI Policy and Governance Working Group Comment at 3.

245 Balwit, A., & Korinek, A. (2022, May 10). Aligned with whom? Direct and social goals for AI systems. https://www.brookings.edu/articles/aligned-with-whom-direct-and-social-goals-for-ai-systems/.

246 Sartori, L., & Theodorou, A. (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics and Information Technology, 24(4). https://link.springer.com/article/10.1007/s10676-022-09624-3.

247 IBM Comment at 8.

248 AI Accountability Policy Report, National Telecommunications and Information Administration. (2024, March). https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report.

249 Gorwa and Michael Veale, 'Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries' (2024) 16(2) Law Innovation and Technology.

250 National Telecommunications and Information Administration (2024). Artificial Intelligence Accountability Policy Report. https://www.ntia.gov/sites/default/files/publications/ntia-ai-report-final.pdf.

251 See, for example, previous writings from the Clinton administration about the Internet, which noted that "governments should encourage industry self-regulation wherever appropriate and support the efforts of private sector organizations to develop mechanisms to facilitate the successful operation of the Internet." 1997 Global Electronic Commerce Framework. Clintonwhitehouse4.Archives.gov.

252 Rodriguez, S. and Schechner, S. Facebook Parent's Plan to Win AI Race: Give Its Tech Away Free. WSJ. May 19, 2024. https://www.wsj.com/tech/ai/metas-plan-to-win-ai-race-give-its-tech-away-free-4bcc080a.

253 See, e.g., Preparedness. (n.d.). OpenAI. https://openai.com/preparedness/; Anthropic's Responsible Scaling Policy. (2023, September 19). Anthropic. https://www.anthropic.com/news/anthropics-responsible-scaling-policy.

254 See, e.g., AI Policy and Governance Working Group Comment at 3-5 ; Public Knowledge Comment at 11-12 ; Hugging Face Comment at 8-9.

255 See, e.g., Stability AI Comment at 4 ("By reducing these costs, open models help to ensure the economic benefits of AI accrue to a broad community of developers and small businesses, not just Big Tech firms with deep pockets.).

256 See, e.g., AI Policy and Governance Working Group at 1 ("Openly available data, code, and infrastructure have been critical to the advancement of science, technological innovation, economic growth, and democratic governance. These open resources have been built and shared in the context of commitments to open science, to expanding industry and markets, and to the principle that some technologies should be widely available for maximum public benefit, while allowing for control of access to data, code, and infrastructure as necessary for safety and security purposes.").

257 Nix, N., Zakrzewski, C., De Vynck, G., (2024, March 10) Silicon Valley is pricing academics out of AI research. Washington Post. https://www.washingtonpost.com/technology/2024/03/10/big-tech-companies-ai-research/.

258 Affiliation of research teams building notable AI systems, by year of publication. (n.d.). Our World in Data. https://ourworldindata.org/grapher/affiliation-researchers-building-artificial-intelligence-systems-all.

259 Rand Comment at 3 ("Publishing foundation model weights can aid in AI safety research.").

260 Google Comment at 2 ("While the benefits of open AI models are profound, there is also a risk that their use accelerates harms, like deepfake imagery, disinformation, and malicious services.").

261 Consistent with this recommendation, the federal government has taken several significant steps toward collecting a more high-quality evidence base. For example, Section 4.2(a) of Executive Order 14110 provides for the collection of information by the federal government from developers of certain dual-use foundation models, including certain "results of any developed dual-use foundation model's performance in relevant AI red-team testing[.]"

262 In our report on AI accountability policy, we stressed the importance of independent audits and assessments in lieu of sole reliance on internal self-assessments. See https://www.ntia.gov/sites/default/files/2024-04/ntia-ai-report-print.pdf at 20-21, 46-49. We noted that "[d]eveloping regulatory requirements for independent evaluations, where warranted, provides a check on false claims and risky AI, and incentivizes stronger evaluation systems." Id. at 48. We concluded that "[i]ndependent AI audits and evaluations are central to any accountability structure[]" and "[t]here are strong arguments for sectoral regulation of AI systems in the United States and for mandatory audits of AI systems deemed to present a high risk of harming rights or safety – according to holistic assessments tailored to deployment and use contexts." Id. at 70, 73. We recommended that the federal government "work with stakeholders as appropriate to create guidelines for AI audits and auditors[]" and "require as needed independent evaluations and regulatory inspections of high-risk AI model classes and systems." Id. at 70, 73.

263 U.S. Copyright Office. Copyright and Artificial Intelligence. https://www.copyright.gov/ai/.

264 Department of Energy. Frontiers in Artificial Intelligence for Science, Security and Technology https://www.energy.gov/fasst.

[265] Model weight restrictions based on non-expressive activity – for example, on a model's demonstrated capability to evade human control – would face fewer legal challenges than restrictions based on expressive activity. However, courts would need to determine which, if any, model weight restrictions counted as expressive speech.

[266] See, e.g., Narayanan, A., & Kapoor, S. (March 21, 2024). AI safety is not a model property. AI Snake Oil. https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property.

[267] Artificial intelligence: Performance on knowledge tests vs. Number of parameters. (2023). https://ourworldindata.org/grapher/ai-performance-knowledge-tests-vs-parameters. Owen, D. (2023, June 9). How Predictable Is Language Model Benchmark Performance? Epoch AI. https://epochai.org/blog/how-predictable-is-language-model-benchmark-performance.

[268] The terms "8B" and "70B" mean the models have 8 billion and 70 billion parameters, respectively.

[269] Saplin, M. (2024, April 18). Llama 3 8B is better than Llama 2 70B. Dev.To. https://dev.to/maximsaplin/llama-3-in-8b-and-70b-sizes-is-out-58pk.

[270] Roser, M., et al. (2023, March 28). What is Moore's Law? Exponential growth is at the heart of the rapid increase of computing capabilities. Ourworld, https://ourworldindata.org/moores-law.

[271] Depending on the policy options under consideration, these developments could counsel in favor of either "broader" or "narrower" thresholds for inclusion. For example, continued decreases in the cost of training powerful models could weigh in favor of more narrowly defining the set of models subject to certain requirements, because those requirements could become impossible to effectively enforce if a very large group of people are each capable of training those models. See also Lambert, N. Interconnects DBRX: The new best open model and Databricks' ML strategy. www.interconnects.ai. https://www.interconnects.ai/p/databricks-dbrx-open-llm (describing "Mosaic's Law", a phenomenon coined by the former CEO of Mosaic whereby training "a model of a certain capability will require 1/4 the [money] every year" due to technological advances). On the other hand, decreases in the amount of model parameters or training compute necessary to achieve a certain level of capability could weigh in favor of more broadly defining the technical characteristics that would subject a model to policy requirements, because models with the same technical characteristics may increase in their capabilities – and therefore risks – over time.

[272] Evo: DNA foundation modeling from molecular to genome scale. Arc Institute. (2024, February 27). https://arcinstitute.org/news/blog/evo. Note, however, that the Evo model contains approximately 7 billion parameters, fewer than the tens-of-billions threshold set forth in the EO.

[273] Hong, W., et al. (2022). CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. ArXiv. https://arxiv.org/pdf/2205.15868; Stable Diffusion 3. (February 22, 2024). Stability AI. https://stability.ai/news/stable-diffusion-3.

[274] Create Realistic Deepfakes with DeepFaceLab 2.0. (2023, November 16). https://www.toolify.ai/ai-news/create-realistic-deepfakes-with-deepfacelab-20-45020.

[275] See Pan Alexander, et al. (2023) Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v202/pan23a.html.

[276] Vadapalli, Sreya et al. (May 2022) Artificial Intelligence and machine learning approaches using gene expression and variant data for personalized medicine. National Institutes of Health National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10233311/.

[277] Selinger, Evan. (November 13, 2021) Facebook's next privacy nightmare will be a sight to see. Boston Globe. https://www.bostonglobe.com/2021/11/12/opinion/facebooks-next-privacy-nightmare-will-be-sight-see/.

[278] Jerome, J. (2020, September 26). The Race to Map Reality So Silicon Valley Can Augment It Is On. https://thewire.in/tech/silicon-valley-augmented-reality-facebook-google.

[279] See Metz, Cade. (May 13, 2024) OpenAI Unveils New ChatGPT That Listens, Looks, and Talks. The New York Times. https://www.nytimes.com/2024/05/13/technology/openai-chatgpt-app.html.

[280] The U.S. Government is currently defining AI and AI-enabling talent in OMB M-24-10 as: individuals with positions and major duties whose contributions are important for successful and responsible AI outcomes. AI and AI-Enabling Roles include both technical and non-technical roles, such as data scientists, software engineers, data engineers, data governance specialists, statisticians, machine learning engineers, applied scientists, designers, economists, operations researchers, product managers, policy analysts, program managers, behavioral and social scientists, customer experience strategists, human resource specialists, contracting officials, managers, and attorneys.

[281] Rand Comment at 4 ("The most common approach to structured access is to create flexible application programming interfaces (APIs) that allow researchers, small businesses, or the public to access the model.").

[282] Shevlane, T. (2022). Structured Access: An Emerging Paradigm for Safe AI Deployment. University of Oxford. at 20. https://arxiv.org/abs/2201.05159.

[283] See, e.g., CDT Comment at 39-40 (discussing potential First Amendment considerations that could be implicated in the regulation of open foundation models and their weights).

[284] Bucknall, B., & Trager, R. (2023). Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements. Oxford Martin School. https://www.oxfordmartin.ox.ac.uk:8443/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements.

[285] See, e.g., Narayanan, A., & Kapoor, S. (March 21, 2024). AI safety is not a model property. AI Snake Oil. https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property.

[286] Chopra, R., et al. (2023). Joint Statement on Enforcement Efforts against Discrimination and Bias in Automated Systems. https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf.

About NTIA

The National Telecommunications and Information Administration (NTIA), located within the Department of Commerce, is the Executive Branch agency that is principally responsible by law for advising the President on telecommunications and information policy issues. NTIA's programs and policymaking focus largely on expanding broadband Internet access and adoption in America, expanding the use of spectrum by all users, and ensuring that the Internet remains an engine for continued innovation and economic growth. These goals are critical to America's competitiveness in the 21st century global economy and to addressing many of the nation's most pressing needs, such as improving education, health care, and public safety.

For more information, please visit us at **ntia.gov**

The National Telecommunications and Information Administration
Herbert C. Hoover Building (HCHB)
U.S. Department of Commerce
National Telecommunications and Information Administration
1401 Constitution Avenue, N.W.
Washington, D.C. 20230