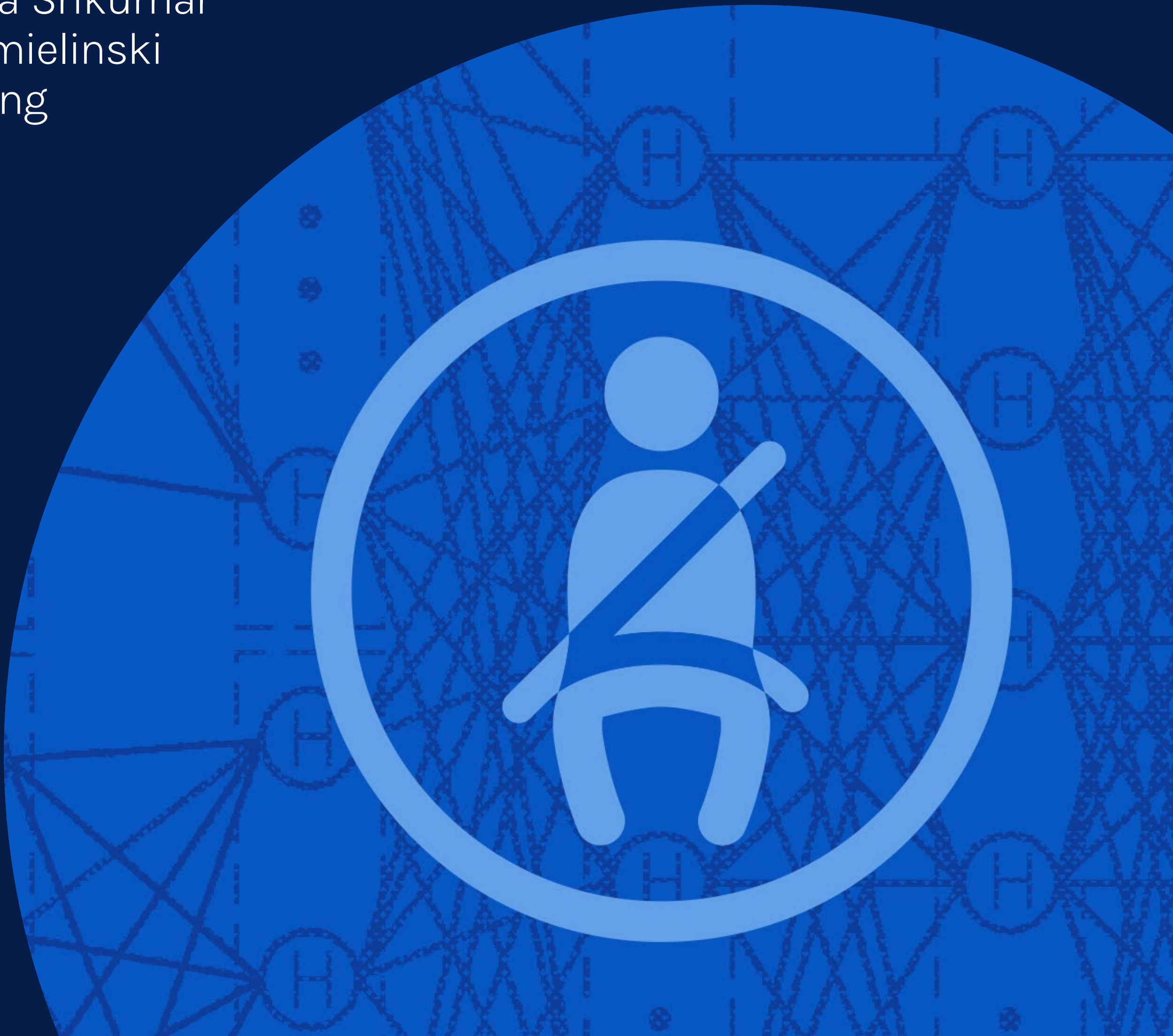
PARTNERSHIP ON AI

# Risk Mitigation Strategies for the Open Foundation Model Value Chain

## Insights from PAI Workshop Co-hosted with GitHub

Madhulika Srikumar
Kasia Chmielinski
Jiyoo Chang



# Contents

# Introduction

Policymakers worldwide are designing interventions to ensure AI development is safe and responsible. In the EU, the newly formed AI Office is drafting codes of practice for general purpose AI models, while in the US, state legislatures are considering laws about who's liable for downstream changes made to AI models. These efforts remind us why we need to track the rapidly changing AI value chain. **As the AI ecosystem grows more complex, with more actors and shifting roles, it's crucial to understand this intricate web to craft effective policy.**

These policy discussions not only recognize the potential risks associated with the modification and use of AI models by actors along the value chain, but also raise questions about the responsibilities of model providers beyond the initial development phase. Such questions are particularly complex for open foundation models,[A] which can be freely accessed and modified by others.

**To address these questions, PAI and GitHub jointly hosted an in-person workshop** on April 11, 2024, convening experts across industry, academia, and civil society to explore safeguards for state-of-the-art[B] open foundation models and roles and responsibilities within the value chain.

Understanding the generative AI value chain[C] is crucial for developing effective strategies to govern open models. Mapping out the various stages of the AI value chain helps us pinpoint where interventions can make a difference, and what risk mitigation strategies various actors can enact. This understanding can also help us explore new approaches for releasing future cutting-edge models, such as staged and component releases, to balance the benefits of openness with the need for responsible use and monitoring.

> Mapping out the various stages of the AI value chain helps us pinpoint where interventions can make a difference, and what risk mitigation strategies various actors can enact.

[A] Foundation models are large-scale base models trained on vast amounts of data, capable of being adapted to a wide range of downstream tasks through methods like fine-tuning or prompting. These models, also known as 'general purpose AI', serve as starting points for developing more specialized AI systems across scientific and commercial domains. Increasingly, these models are being integrated into operating systems and services as AI assistants or "agents", capable of understanding personal context and eventually performing complex tasks across applications.

[B] Current state-of-the-art models can generate synthetic content like text, images, audio, and video. They may be narrow-purpose, focusing on specific tasks, modalities, or data types, like those trained on biological sequences, or general-purpose.

[C] While some literature uses the term 'supply chain' in this context, this document uses 'value chain' to align with terminology in the EU AI Act. The Act appears to use 'value chain' to refer to the entire ecosystem of actors involved in bringing an AI system to market and maintaining it. This usage is closer to a traditional supply chain concept, focusing on the network of actors involved in producing and delivering AI systems, rather than specifically on how each actor adds value. We adopt this terminology for consistency with emerging regulatory frameworks.

## What is an Open Foundation Model?

Open foundation models refer to AI models whose "building blocks," most notably model weights, are released openly, allowing others to use, study, modify, and build upon them.[D] While open models offer significant benefits, such as increased accountability, innovation, competition, and enabling critical safety research, their open nature can make it more challenging to assess and mitigate risks.[1]

Concerns arise when model weights are released openly because downstream developers can fine-tune[E] the models and circumvent safety guardrails put in place by the original developers.[2] As models become increasingly capable of generating realistic content and enabling execution of complex tasks, the chances for malicious use heightens. Bad actors can misuse open models through fine-tuning to create harmful content, including fake or manipulated imagery for harassment, fraud or disinformation. Additionally, because many open models can be run locally on relatively inexpensive hardware, they increase the chances for misuse at a low cost in ways that may not be directly monitored.

In contrast to open foundation models, more closed models[F] typically restrict access to their weights. They are accessed through APIs, allowing the model providers to control and monitor their usage. However, they are not immune to misuse or reckless use, as downstream developers can still prompt these models and, in some cases, fine-tune them, bypassing safety measures. The key difference is that more closed models offer providers more direct levers to monitor and moderate usage, potentially allowing for quicker responses to misuse or reckless use. In October 2023, PAI published its Guidance for Safe Foundation Model Deployment, which included specific recommendations for closed and restricted releases.

> As models become increasingly capable of generating realistic content and enabling execution of complex tasks, the chances for malicious use heightens.

[D] While 'openly released' suggests unrestricted access, SOTA open foundation models typically have license restrictions based on factors such as user base size, age, or geographical location. Additionally, 'openness' in this context refers to the release of model weights, though some open FMs make more components available. This definition is consistent with the recent US Executive Order's notion of "foundation models with widely available model weights." It differs from broader definitions of open source AI, such as those being developed by the Open Source Initiative, which may require openness of more components (e.g., training data, code). For example, AI2's OLMo and Eleuther's Pythia are considered more open than Meta's Llama 3 due to fewer restrictions on access and use and more open components.

[E] Fine-tuning is the process of adapting a pre-trained model to a specific task or domain by training it on additional data. This process requires technical expertise and resources, and can make models either safer or less safe depending on the intent and implementation.

[F] The openness of AI models exists on a spectrum, with 'closed' models ranging from API-only access to fully proprietary systems. Open models can range from those with partially released components to fully open-source models with all weights and training data available. This spectrum is particularly notable in foundation models, where the release of model weights is a significant factor in determining openness since the weights allow for deeper modification. In contrast, other AI models like recommender systems tend to be more consistently closed.

## The Open Foundation Model Value Chain

Our mapping of the AI ecosystem has led to a framework that captures the current state of the open foundation model value chain. This framework illustrates the complex web of actors and processes involved in developing and deploying these models.

**Figure 1: Value Chain for Open Foundation Model Governance**

MLOPS AND EVALUATION PROVIDERS

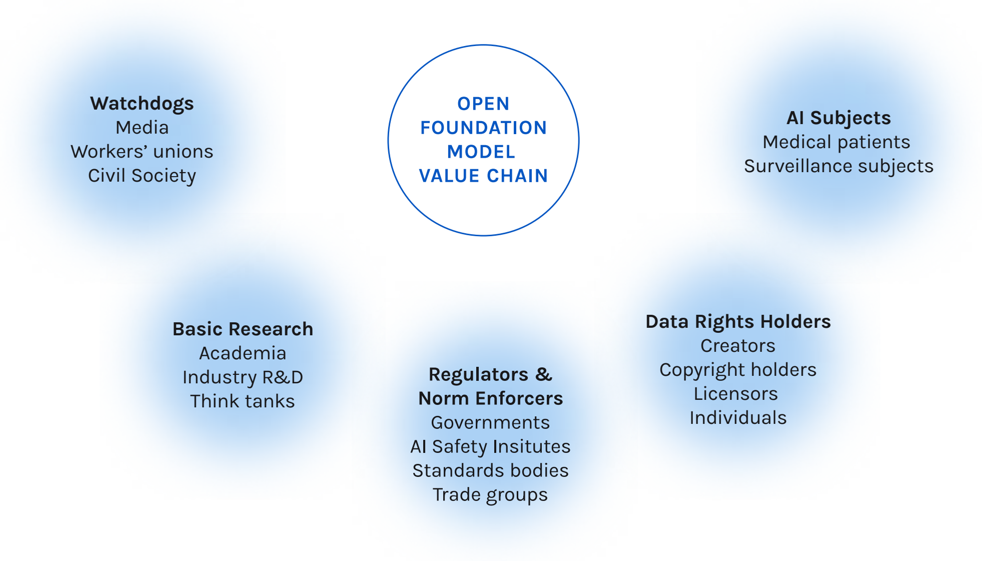
MITIGATIONS APPLY

| ACTOR | DESCRIPTION | EXAMPLES |
|---|---|---|
| COMPUTE AND CLOUD PROVIDERS | Provide the necessary computing power and infrastructure to train and run foundation models. Users can either build their own computing clusters using hardware from compute providers or leverage the cloud infrastructure and services offered by cloud providers. | Nvidia and AMD *(provides hardware)*<br>AWS<br>Google Cloud<br>Microsoft Azure<br>Coreweave<br>IBM Cloud |
| DATA PROVIDERS | Provide datasets for model training and refinement, whether created intentionally for this purpose or not | Common Crawl *(non-AI specific)*<br>LAION *(open image dataset)*<br>Eleuther's Pile *(open text dataset)*<br>Scale AI *(services)* |
| OPEN FOUNDATION MODEL PROVIDERS | Train and distribute model weights and other components, that may be intended for research purposes, open source development (allowing free access, modification, and distribution), or commercial provision (with potential for paid services or products related to the model) | Meta's Llama 3<br>Google's Gemma 2<br>AI2's OLMo<br>Stability AI's Stable Diffusion<br>EleutherAI's Pythia<br>Mistral<br>IBM's Granite code models |
| MODEL HUBS & HOSTING SERVICES | Make foundation models, including adapted foundation models, and other components available to downstream developers. Hosting services may include solutions that enable running inference for end-users | GitHub *(shares components)*<br>Civitai *(enables inference)*<br>Hugging Face *(does both)* |
| MODEL ADAPTERS & OPTIMIZERS | Fine-tune or otherwise adapt an open foundation model for a specific purpose or domain, entities that optimize open foundation models to work on diverse hardware, and service providers that support these modifications | LMSYS's Vicuna *(fine-tuned Llama)*<br>Meditron *(fine-tuned Llama for medical domain)*<br>Databricks *(supports modification)*<br>Scale AI *(supports modification)* |
| APP DEVELOPERS, SERVICE DEVELOPERS, MODEL INTEGRATORS | Integrate open foundation models into AI systems, potentially with additional safety and capability enhancements. Deploy these integrated systems as: a) Services for developers (e.g., hosted models accessible via API) b) Applications or services for end-users (e.g., chatbots, digital assistants including OS-level integrations, creative tools) | Meta AI's web UI<br>Midjourney |
| DISTRIBUTION PLATFORMS | Provide platforms such as social media networks, app stores or content-sharing sites, where applications and AI-generated content may be shared | Facebook<br>Apple App Store<br>OpenAI GPT Store<br>YouTube |
| USERS | Use FM-based apps and services, either as enterprise (B2B) or consumer (B2C) | Finance companies<br>Chatbot users<br>Medical patients |

OCCURS AT MULTIPLE STAGES OF THE VALUE CHAIN

| ACTOR | DESCRIPTION | EXAMPLES |
|---|---|---|
| ML OPS & EVALUATION PROVIDERS | Offer tools for efficient ML lifecycle management, performance evaluation, safety assessment, continuous improvement, and auditing | Hugging Face's OpenLLM Leaderboard<br>Stanford's HELM<br>MLCommons<br>METR<br>Independent auditors |

Figure 2: Actors outside the value chain



While Model Providers are typically the focus, other important players like Model Adapters, Hosting Services, and Application Developers are often overlooked, even though they play key roles in facilitating discovery of models, enabling their widespread use, as well as preventing or reducing potential harms from both open and closed foundation models. For instance, in addressing the risk of non-consensual intimate imagery (NCII), Model Providers can ensure training data is sourced responsibly and screened for harmful content, Model Providers and Model Adapters can implement content filters to avoid generating NCII, Hosting Services can enforce content moderation policies on models tailored to create NCII, and Application Developers can integrate detection and user reporting mechanisms. Further downstream, social media platforms where NCII might be distributed can take steps to limit its spread. It's important to note that these interventions don't always occur as a single chain of events. The specific levers available to each actor to limit NCII generated using open foundation models depends on the nature of the incident, and we've seen real-world incidents involving their misuse.[3] While closed models allow for more direct enforcement of safeguards, open models require a distributed value chain approach to prevention and response.

That said, the AI value chain is complex, with various actors having different incentives, levels of awareness, and control when it comes to responsible AI development. Studies suggest that this complexity can lead to responsibility falling through the cracks, resulting in potential misuses, unintended uses, or other challenges going unaddressed.[4, 5] Insufficient documentation and guidance for downstream developers and users can sometimes further add to the complexity. To tackle these challenges, mechanisms for communication and collaboration among value chain actors is essential, which can help inform the development and implementation of effective risk mitigation strategies outlined below.

Outside actors like journalists and governments also have a role to play. They can investigate AI's impact on society and drive regulatory reforms that thoughtfully contend with the evolving set of actors and capabilities in this nascent economy. They can create pressure to encourage responsible practices that balance openness with effective risk mitigation strategies. As the state of the art in foundation models advances, policymakers and stakeholders need to keep up with how these models are integrated into the real world, including the handoffs between different actors, to manage their impacts.

## Key Considerations

Our analysis of the AI value chain and current governance landscape has identified the following key considerations:

- **The value chain is not linear.** While it is conceptually useful to understand the value chain as a linear progression, many functions occur in parallel, non-linearly, or repeatedly based on the use case. For example, fine-tuning may or may not occur, and different types of data (size, annotated, etc.) are required at various points. Governance may need to be adjusted to take into consideration not just the function of the actor but also the point at which the action occurs.
- **Governance interventions should distinguish between the various intentions of actors.** Product-safety[G]-oriented policies should also focus on preventing reckless use[6] of foundation models (stemming from misguided or uninformed actions) rather than solely misuse (intentional malfeasance, such as removing safety guardrails from open foundation models). Misuse considerations may require policies and investments focused on public safety[H] and societal preparedness.
- **Certain actors in the value chain may have similar governance capabilities and affordances**, even if they operate at different layers. For example, model providers and model hosting services that allow users to run inference directly on their platforms may have governance capabilities similar to those of application developers like usage monitoring.
- **Many layers of the value chain are evolving rapidly**, especially the more nascent ones like the Model Adapter and App Developer Layers. We expect to see further differentiation or possibly consolidation over time. As the Model Adapter layer matures, it may warrant closer policy attention, as modifications to base models could potentially introduce new risks.
- **Several actors in the ecosystem span multiple layers of the value chain**, particularly in compute, fine-tuning, tooling, and deployment. There is also a trend towards single actors offering fully vertically integrated services, from compute to end-user distribution platforms, for closed AI models. A single actor may thus be responsible for many forms of governance when that actor operates in multiple layers. In those cases, the closed model paradigm consolidates the roles and responsibilities of multiple actors in the open model value chain, potentially simplifying governance and accountability structures but also limiting visibility and centralizing control over the model development and deployment process.

[G] Product safety focuses on ensuring that AI models and applications are designed and deployed in a way that minimizes risks to individual users, such as protecting privacy, preventing unintended harm, and maintaining transparency.

[H] Public safety extends beyond model and system-level mitigations, encompassing measures such as biosecurity protocols to prevent AI misuse in developing biological weapons media literacy programs to counter AI-generated disinformation.

## Risk Mitigation Strategies

To effectively address the challenges associated with open foundation models, it's crucial to consider a range of risk mitigation strategies. These strategies can be categorized into three high-level groups:

**PREVENT**

**DETECT**

**RESPOND**

While certain actors may be better positioned to implement specific strategies, all actors in the AI value chain should explore ways to contribute to these efforts within their roles and contexts.

**Figure 3: Risk Mitigations in the Open Paradigm: Initial Workshop Consensus**

● Primary actor responsible
? Merits exploration

| Mitigation | Model Providers | Model Adapters | Model Hosting Services | App Developers |
|---|---|---|---|---|
| **PREVENT** | | | | |
| Responsibly Source and Filter Training Data | ● | ? | | |
| Conduct Internal And External Safety and Misuse Evaluations | ● | ? | | ? |
| Implement Disclosure Mechanisms for Ai-Generated Content | ● | | | ● |
| Provide Downstream Use Guidance and Tooling | ● | ? | ? | |
| Publish a Responsible AI License | ● | | | |
| Establish Clear and Consistent Content Moderation for Hosted Models | | | ● | |
| Implement Use Case-Specific Safety Measures | | ● | | ● |
| Implement Staged Release and Phased Deployments | ● | | | |
| Develop and Implement Durable Model-Level Safeguards | ● | ? | | |
| Release Models with Digital Signatures or 'Fingerprints' | ● | | | |
| **DETECT** | | | | |
| Monitor Misuses, Unintended Uses, and User Feedback | ● | ? | ? | ● |
| Implement Incident Reporting Channels | ● | ? | ● | ● |
| **RESPOND** | | | | |
| Enforce Consequences for Policy Violations | | | ● | ● |
| Establish Decommissioning and Incident Response Policies | ● | ? | ● | ● |
| Develop and Adhere to Transparency Reporting Standards | ● | | ● | ● |

### PREVENT

- Proactive technical and policy measures[I] to support responsible use, and anticipate and reduce the likelihood of misuse or unintended consequences before model deployment.
- Key strategies include performing internal and external safety and misuse evaluations, providing downstream use guidance and tooling, and implementing disclosure mechanisms for AI-generated content.

### DETECT

- Technical and policy interventions to identify instances of misuse or unintended consequences post-deployment.
- Key strategies include monitoring misuses and unintended uses, encouraging user feedback, as well as implementing incident reporting channels.

### RESPOND

- Actions taken to address identified instances of misuse or unintended consequences and prevent future occurrences.
- Key strategies include enforcing consequences for policy violations, establishing decommissioning and incident response policies, and developing transparency reporting standards.

See Appendix A for a comprehensive overview of the 15 risk mitigation strategies. Many of these draw from the PAI Guidance for Safe Foundation Model Deployment. They cover each strategy, the actors[J] who could implement them, and the potential challenges involved.

## Challenges to Implementation

While these risk mitigation strategies provide a starting point, their implementation faces three key challenges: technical limitations, complex content decisions, and the varying effectiveness of different interventions.

1. **Technical limitations**
   Technical safeguards face significant hurdles in implementation and durability. For example, creating synthetic media disclosures (such as watermarks) that are resilient to removal or modification is a complex technical challenge. These safeguards need to persist across various transformations and uses of the content, which is particularly difficult in open ecosystems where downstream modifications are common.
2. **Complex content decisions**
   Even model-level mitigations involve nuanced content-related choices. For instance, determining appropriate content filters for large language models involves nuanced decisions about what types of outputs should be restricted. In a prompting paradigm, this might involve deciding which topics or types of language should trigger refusals, balancing safety concerns with the model's utility. These decisions become even more challenging in open ecosystems, where upstream choices can have wide-ranging and sometimes unforeseen impacts on downstream applications.
3. **The varying effectiveness of different interventions**
   The effectiveness of different mitigation strategies can vary widely. Some interventions, like responsible AI licenses, while well-intentioned, may have limited

[I] Policy measures are non-technical interventions for governing AI systems, including transparency (e.g., disclosures), accountability (e.g., audits), and content governance (e.g., moderation guidelines). These differ from technical measures, which involve modifications to the AI system, like filters to prevent harmful content generation.

[J] There is a primary and secondary actor listed for each mitigation below. Secondary actor can be an entity in the AI value chain that should consider, adapt, or support the implementation of a risk mitigation strategy.

practical impact due to enforcement challenges. Other approaches, such as external evaluations or pre-release audits, require a mature ecosystem of skilled actors to be truly effective. Further, when considering risk management strategies, it's crucial to evaluate the net impact of open model releases. This includes assessing how open models might improve defenses against potential misuse, potentially outweighing risks in certain areas.

**Key questions remain:**

- Which actors are best suited to implement particular mitigations in the context of open foundation models?
- What additional guidance or case studies of value chain approaches would be useful?

## Future-Proofing AI Governance

As we develop and refine these risk mitigation strategies, it's crucial to consider the future implications of more capable and widely adopted open foundation models.

As we develop and refine these risk mitigation strategies, it's crucial to consider the future implications of more capable and widely adopted open foundation models.

By understanding the roles, capabilities, and interactions of different actors in the AI value chain, we can explore more nuanced approaches for sharing foundation models and their components in order to make progress towards industry best practice.

### Staged releases

Model providers can design release strategies that gradually expand access to a model over time, based on evidence of safety and efficacy. This could involve initially releasing a model to a small group of trusted partners, then to a wider community of researchers and developers, and eventually to the general public. By monitoring and assessing the impacts of the model at each stage, through restricted access, providers can make informed decisions about when and how to expand access. A value-chain perspective helps identify the appropriate actors to involve at each stage, ensuring that the model is thoroughly tested and validated by relevant stakeholders before wider release. While the specific criteria for which models should undergo such staged releases are still being debated, this approach may be valuable for models with significant capability advancements or novel modalities.

### Component-level releases

Providers can share specific components in addition to or in place of model weights, such as training data, model architectures, safety evaluation and tuning tools, or output layers. This approach can promote transparency and collaboration without necessarily requiring full model disclosure. For example, releasing a model's architecture and training data can support researchers in auditing the model for potential biases or vulnerabilities, while still maintaining safeguards based on the model's intended use and potential risks. By considering the roles and capabilities of different actors in the value chain, providers can determine which components to release to specific stakeholders to maximize the benefits of openness while mitigating risks.[7]

## Next Steps for Open Foundation Model Governance

Safety must be a core consideration, and it cannot be thoroughly addressed by only scrutinizing AI models. Models are integrated into AI systems, which in turn are deployed in specific contexts and environments. Understanding the system, in contrast to the model, and its context is vital. The value-chain perspective is valuable in centering the range of actors involved in the processes of building complete AI systems and selecting deployment contexts.

Understanding how AI models may be combined together or might interact with each other and create new governance challenges beyond the scope of a single model or value chain is essential for getting AI governance right.

To support the responsible open release of future cutting-edge models, a research agenda could focus on developing and validating model safeguards and advancing societal readiness.[K] This could involve advancing research and engineering challenges that are currently unfeasible, as well as assessing the coordination and support necessary across the ecosystem to ensure responsible open release.[L] While the science and practice of pre-release evaluations are nascent, there is a need to identify categories of models warranting additional evaluation, the types of evaluations required, and the specific outcomes that signal severe risk thresholds, which should give model providers pause in releasing. Today, risk thresholds that may weigh against open release of models are not yet articulated,[M] though developing shared risk thresholds for frontier models has been identified as a priority for industry and government at the 2024 AI Seoul Summit.[8]

The landscape of open foundation models spans from those with better understood capabilities to those pushing frontier boundaries, necessitating a nuanced approach to openness and risk mitigation. Pursuing risk mitigations strategies for state-of-the-art open foundation models is essential, while striving to maintain their accessibility. We can learn from other fields that have advanced safety measures while preserving access to the underlying technology and the benefits it provides.[N] With the rapidly evolving nature of the technology and associated risks, we must also establish mechanisms to effectively evaluate and adapt risk mitigation strategies over time. This requires ongoing ecosystem monitoring by stakeholders, evidence-based risk assessments,[9] and a commitment to ensuring that risk mitigation efforts remain proportionate and effective.

To achieve these goals and maximize the benefits of openness, all stakeholders should actively engage with the value chain to support informed and responsible use of open foundation models. This involves fostering collaboration among providers, developers, policymakers, and other relevant actors to share knowledge, data, and insights. It also requires careful consideration of trade-offs and stakeholder input when crafting AI governance frameworks to ensure they are well-informed, balanced, and adaptable to the evolving landscape.

[K] Public safety and societal preparedness measures can include security protocols at biological labs to prevent AI misuse in developing weapons, as well as strategies to verify the authenticity of human-generated content and actions in an environment that can become populated by AI-generated content and advanced digital assistants or agents.

[L] Technical research priorities could include developing resilient synthetic media disclosures that resist removal or modification. Support for collaboration could involve shared industry approaches to technical challenges and joint industry-academia initiatives for developing safeguards, potentially including model access and compute resources for research purposes.

[M] The PAI Model Deployment Guidance suggests capability thresholds in addition to compute, recommending that "frontier and paradigm-shifting models" demonstrating self-learning capabilities exceeding current AI or enabling direct real-world actions (agentic systems) should initially be released through staged rollouts and restricted access to establish confidence in risk management before considering open availability.

[N] The debate around end-to-end encryption (E2EE) in messaging services offers a relevant analogy. While critics argued E2EE hinders abuse detection, studies show effective mitigation doesn't always require access to user content. Similarly, balancing open access to AI models with safety measures can draw lessons from this experience, aiming to preserve benefits while addressing risks.

## Acknowledgements

We appreciate the invaluable inputs provided by experts who attended the workshop and those who reviewed this article, including:

- Medha Bankhwal, McKinsey & Company
- Anthony Barrett, UC Berkeley CLTC
- William Bartholomew, Microsoft
- Rishi Bommasani, Stanford CRFM
- Ian Eisenberg, Credo AI
- Marzieh Fadaee, Cohere
- Heather Frase, verAItech
- David Evan Harris, UC Berkeley
- Divyansh Kaushik, Beacon Global Strategies
- Kevin Klyman, Stanford HAI
- Alex Kessler, Microsoft
- Yolanda Lannquist, The Future Society
- Kyle Lo, Allen Institute for AI
- Nik Marda, Mozilla
- Aviv Ovadya, AI & Democracy Foundation
- Rebecca Portnoff, Thorn
- Hadrien Pouget, Carnegie Endowment for International Peace
- Saishruthi Swaminathan, IBM
- Margaret Tucker, GitHub
- Cody Venzke, ACLU
- Rebecca Weiss, MLCommons
- Matt White, Linux Foundation
- Kevin Xu, GitHub

Special thanks to:

- Peter Cihon, GitHub
- Mike Linksvayer, GitHub
- David Gray Widder, Digital Life Initiative at Cornell Tech
- Aviya Skowron, EleutherAI
- Michael Veale, University College London

APPENDIX A

# Detailed Risk Mitigation Strategies

There is a primary and secondary actor listed for each mitigation below. Secondary actor can be an entity in the AI value chain that should consider, adapt, or support the implementation of a risk mitigation strategy.

## PREVENT

### Responsibly Source and Filter Training Data

ACTOR: MODEL PROVIDERS
SECONDARY ACTOR: MODEL ADAPTERS

Model providers should carefully curate and filter their training data to mitigate the risks of misuse by malicious actors and unintended consequences by downstream developers. This involves implementing robust processes to identify and remove potentially harmful content, such as hate speech, explicit material, personally identifiable information (PII), or content that violates intellectual property rights. Providers should also strive to ensure that their training data is diverse, representative, and free from biases that could lead to discriminatory outputs.

One critical example of this mitigation is detecting, removing, and reporting child sexual abuse material (CSAM) from training data. Providers should avoid using data with a known risk of containing CSAM and implement tools and processes to proactively identify and remove any instances of CSAM or related content. This can include using hash-matching techniques to compare training data against known CSAM databases and collaborating with organizations like the National Center for Missing and Exploited Children (NCMEC) to report any identified CSAM. Providers should also take steps to separate depictions or representations of children from adult sexual content in their training datasets to further mitigate the risk of creating models that could be used to generate CSAM.

However, responsibly sourcing and filtering training data can be challenging, particularly for large-scale datasets. It requires significant resources and expertise to develop and maintain effective content moderation processes. The constantly evolving nature of online content and the potential for adversarial attacks, such as data poisoning, can make it difficult to ensure that all harmful content is identified and removed. Balancing the need for diverse and representative data with the imperative to filter out harmful content can also be complex, requiring careful consideration of ethical and societal implications, including responsible handling of demographic data. Additionally, model providers should consider making their training data available for research, scrutiny, and auditing, as well as disclosing their data sources, to promote transparency and enable independent verification of data practices.

### Conduct Internal and External Safety and Misuse Evaluations

ACTOR: MODEL PROVIDERS
SECONDARY ACTORS: MODEL ADAPTERS, APP DEVELOPERS

Perform Internal Safety and Misuse Evaluations: Model providers can conduct internal evaluations of their models prior to release to assess and mitigate potential misuse risks. This can include using pre-release red teaming methods to assess the potential for implemented safety guardrails to be circumvented post-release. For open foundation models, providers may need to focus on hardening the model against specific misuses (e.g., via reinforcement learning from human feedback (RLHF) or reinforcement learning from AI feedback (RLAIF) training) and finding ways to make the model resilient to attempts to fine-tune it onto a dataset that would enable misuse. Other mitigations suggested include providers should "use a high evaluation bar" and hold open models to "a higher bar for evaluating risk of abuse or harm than proprietary models, given the more limited set of post-deployment mitigations currently available for open models." These evaluations can involve fine-tuning a base model to maximize its propensity to perform undesirable actions. Conducting internal safety and misuse evaluations, particularly red teaming exercises, can be resource-intensive and may not fully anticipate all possible misuse scenarios. The rapidly evolving landscape of open foundation models can make it challenging to keep pace with new risks and vulnerabilities.

Conduct External Safety Evaluations: Model providers, including model adapters, can complement internal testing by providing controlled access to their models for third-party researchers to assess and mitigate potential misuse risks. This can include consulting independent parties to audit models using prevailing best practices, identifying potential misuse risks, adapting deployment plans accordingly, and maintaining documentation of evaluation methods, results, limitations, and steps taken to address issues. Enabling robust third-party auditing remains an open challenge requiring ongoing research and attention. External safety assessments like red-teaming, while valuable, may expose models to additional risks if not carefully managed. Balancing the benefits of external input with the potential risks requires thoughtful consideration.

## Implement Disclosure Mechanisms for AI-generated Content

ACTORS: MODEL PROVIDERS, APPLICATION DEVELOPERS

Model providers can embed watermarks or other indirect disclosures into the model's outputs to help trace the source of misuse or harmful content. It has been suggested that model providers use *maximally indelible* watermarks, which are as difficult to remove as possible. Application developers should integrate the model with these safeguards and also embed direct disclosures that are viewer or listener facing to indicate that the content is generated by an AI model.

The open nature of these models presents unique challenges that can make it difficult to ensure the effectiveness and enforceability of prevention strategies. The potential for malicious fine-tuning and circumvention of safety features at the model layer can limit their effectiveness, as models can be modified or used in unintended ways post-release.

Currently, embedding watermarks directly into language model weights is not technically feasible. For non-text media (images, audio, video), various indirect disclosure techniques like watermarking and cryptographic provenance show promise, though each has pros and cons. For text outputs, robust methods don't exist for either open or closed models. However, actors serving inference can implement watermarking during generation for closed models. This approach is less effective for open models, as users can circumvent it by running the model without the watermark implemented in the pipeline. An emerging practice is open-sourcing text watermarking technology. However, this approach may have tradeoffs, including potential vulnerability to adversarial attacks.

### Provide Downstream Use Guidance and Tooling

ACTOR: MODEL PROVIDERS
SECONDARY ACTOR: MODEL ADAPTERS

*This practice could be partially extended to more actors like model hubs who can support the visibility of guidance shared by Model Providers/Adapters.*

Model providers can equip downstream developers (Model Adapters & Optimizers, Application Developers) with comprehensive documentation like model cards and guidance needed to build safe and responsible applications using open foundation models. This can include providing documentation covering details such as suggested intended uses, limitations, steps to mitigate misuse risks, and safe development practices when building on open foundation models. Models with greater openness with open source code, documentation, and data can mitigate reckless use by providing better information for model adapters and application developers. Model providers can also offer downstream safety tools and resources, such as Meta's Purple Llama project, which includes Llama Guard — an openly available foundational model to help developers implement content filtering and avoid generating potentially risky outputs in their applications built on open foundation models. Providing comprehensive downstream use guidance necessitates close collaboration with various stakeholders and ongoing continuous updates. The decentralized deployment and limited control over how open models are used can make it difficult to ensure adherence to the provided guidance.

### Publish a Responsible AI License

ACTOR: MODEL PROVIDERS

Model providers can publish a responsible AI license that prohibits the use of open foundation models for harmful applications. The license could clearly define what constitutes harmful use and outline the consequences for violating the terms of the license. Providers can also consider requiring users to agree to the license terms before accessing the model. Enforcing a responsible AI license may be challenging, as open models can be easily shared and used outside the provider's control. Providers may need to rely on legal action or community pressure to hold violators accountable, recognizing the limits of governance by licenses, which can typically only be enforced by the rightsholder or a delegated agent. Mechanisms to fund such enforcement may need to be developed. License terms may be conflicting and subject to different interpretations. Responsible AI Licenses conflict with open source norms that do not restrict use cases when sharing software under open source licenses. This may push users to adopt more open alternatives which may unintentionally lead to decreased use of and investment in the safest models.

### Establish Clear and Consistent Content Moderation for Hosted Models

ACTOR: MODEL HOSTING SERVICES

Model hosting services could establish a structured process for ongoing moderation, including receiving, reviewing, and actioning violations for hosted models. This process could review the documentation and downstream use guidance provided by model providers alongside the AI components. This process could assess whether the model aligns with the hosting service's policies and standards for responsible AI development and deployment, as well as applicable laws. The review process can include:

- Structured reporting forms that support review and response at scale for possible violations, e.g., abuse, private information that poses security risks, intellectual property laws, and other violations of acceptable use policies.

- Evaluation of the completeness and clarity of the model documentation, including information on the training data, model architecture, performance metrics, and known limitations or biases.
- Assessing the adequacy of the downstream use guidance, including recommendations for safe and responsible use, potential misuse risks, and any restrictions or constraints on use.
- Determining whether the model has undergone appropriate testing, evaluation, and risk assessment processes, as evidenced by the documentation.
- Consistent interpretation of model licenses for which hosting services may receive takedown requests. This could involve establishing lists of licenses that hosting services will consider due to their actionable and sufficiently non-vague terms and provisions.

As an alternative or complementary approach for models meeting specific criteria, model hosting services could establish a pre-upload review process for model documentation and downstream use guidance before hosting or distributing models. This proactive review could ensure that models align with the hosting service's policies and standards for responsible AI development and deployment. The review process can include:

- Evaluating the completeness and clarity of the model documentation, including information on the training data, model architecture, performance metrics, and known limitations or biases.
- Assessing the adequacy of the downstream use guidance, including recommendations for safe and responsible use, potential misuse risks, and any restrictions or constraints on use.
- Determining whether the model has undergone appropriate testing, evaluation, and risk assessment processes, as evidenced by the documentation.
- Making the checklist or criteria used in this review process transparent to model providers and the public.

Pre-upload reviews can be challenging for iterative development, which is common in software development. It may also be difficult to apply this process to models developed openly from idea to actual training. Such reviews could potentially turn the hosting service into a publisher rather than a neutral platform, raising additional concerns.

## Implement Use Case-Specific Safety Measures

ACTORS: MODEL ADAPTERS, APPLICATION DEVELOPERS

Model adapters and application developers should implement safety measures tailored to their specific use cases to mitigate potential misuse risks. Examples of use case-specific safety measures that application developers and model adapters can implement include:

- Implementing application-specific content filters and output restrictions to prevent the generation of harmful, inappropriate, or sensitive content.
- Employing techniques like reinforcement learning from human feedback (RLHF) to fine-tune models for specific use cases while mitigating potential misuse risks.
- Conducting ongoing evaluations and de-biasing efforts to ensure the adapted model's outputs remain safe and unbiased for the intended use case.
- Implementing robust monitoring and incident response processes to detect and address any misuse or unintended consequences promptly (more below).

However, developing and maintaining use case-specific safety measures can be resource-intensive, especially for smaller organizations or developers. It may be challenging to anticipate all potential misuse cases or unintended consequences for a given use case.

## Implement Staged Release and Phased Deployments

ACTOR: MODEL PROVIDERS

Model providers could use a staged-release approach, starting with a restricted or structured access release (e.g., behind an API) to monitor for novel risks and harms before proceeding to a full public release of model weights. The PAI Guidance recommends that frontier model providers "initially err towards staged rollouts and restricted access to establish confidence in risk management before considering open availability," if their models demonstrate self-learning capabilities exceeding current AI, enabling execution of commands online or other direct real-world actions (agentic systems). These models may possess unprecedented capabilities and modalities not yet sufficiently tested in use, carrying uncertainties around risks of misuse and societal impacts. Over time, as practices and norms mature, open access may become viable if adequate safeguards are demonstrated. Another approach suggested could be to restrict access to model weights while allowing access to other components to enable researchers and developers to study and build on the model without the risk of uncontrolled proliferation. Access to different components of the models is crucial for realizing benefits but also carries risks. However, implementing staged release and phased deployments is not without challenges. Even with structured access or limited initial release to a smaller group, there is still a risk of model leakage or exfiltration, which could lead to the unintended of model weights.

## Develop and Implement Durable Model-Level Safeguards

ACTOR: MODEL PROVIDERS

SECONDARY ACTORS: MODEL ADAPTERS

Model providers can implement safety features directly into the architectures and interfaces of open foundation models to restrict unsafe uses and mitigate misuse risks. This can include:

- **Content filters:** Model providers can implement filters that detect and block the generation of harmful or inappropriate content, such as hate speech, explicit material, or violent content. Application developers should also integrate these filters with the model in the system and implement additional application-specific filters to detect and block harmful content.
- **Output restrictions:** Model providers can place limits on the types of outputs the model can generate, such as preventing the generation of personal information, financial data, or other sensitive content. Application developers should adhere to these restrictions and implement additional output restrictions tailored to their specific use case.

This responsibility extends to applications built on both open and closed models. The openness of the underlying foundation model likely does not marginally increase the risks of toxicity, bias, or misuse in the resulting applications. Nonetheless, safety features at the application layer are still necessary to mitigate downstream misuses. Additionally, Model Adapters could seek to preserve or augment the safeguards that were created at the model layer by providers.

Model providers should invest in research on methods to pre-train models with difficult-to-remove safety mechanisms, such as self-destructing models that break when users attempt to alter or remove safety guardrails. These safety features should be designed to be difficult to remove or bypass post-release. Research in this area is still in fairly early stages, and more work is needed to develop and test these approaches. The openness of foundation models presents challenges in ensuring the effectiveness and enforceability of these safety features, as models can be modified or used in unintended ways post-release.

### Release Models with Digital Signatures or ‘Fingerprints’’

ACTOR: MODEL PROVIDERS

Model providers can release their models with digital signatures or “fingerprints” to enable greater visibility, traceability, and accountability for use. These digital signatures or fingerprints can help track the provenance of the model and its outputs, making it easier to identify the source of misuse or harmful content. Techniques such as watermarking or embedding unique identifiers into the model’s weights can be used to create these digital signatures. However, the effectiveness of digital signatures or fingerprints in preventing misuse may be limited, as determined adversaries may still find ways to remove or obfuscate these identifiers. Balancing with user privacy concerns and the open nature of the models can be challenging.

## DETECT

### Monitor Misuses, Unintended Uses, and User Feedback

ACTORS: MODEL PROVIDERS, APPLICATION DEVELOPERS
SECONDARY ACTORS: MODEL ADAPTERS, MODEL HOSTING SERVICES

Model providers, hosting services, and application developers could establish monitoring processes to review downstream usage, unintended uses, misuses, and user feedback on their open foundation models and derivative applications. Model providers should monitor public forums, social media, and other channels where their models are being discussed or used to identify potential misuses or unintended consequences. They should also establish clear channels for users to report issues or concerns.

- **Model hosting services** may provide models for download or use via online inference. When a model hosting service provides online inference, intermediaries they have more direct control and visibility over how the model is being used. Online inference platforms therefore should directly monitor the usage of hosted models and enforce their terms of service, which should prohibit harmful or malicious use. For models that are downloaded and run locally or elsewhere, monitoring or reporting by the model hosting service may be infeasible since users can run them on their own devices. In these cases, hosting services should monitor reports of abuse and enforce their terms of service to reduce discovery and use of concerning models, particularly those modified or otherwise pre-configured to do harm.
- **Application developers** should closely monitor user interactions with their applications and promptly address any reports of misuse or unintended consequences. All actors should collaborate and share information about identified issues to help improve the overall safety and responsibility of the open foundation model ecosystem.

However, maintaining processes to review downstream usage requires ongoing resources and may be complicated by the decentralized nature of open models. This challenge is particularly relevant at the model layer, where providers and adapters may have limited visibility into how their models are being used once they are openly available. Balancing the level of monitoring with user privacy concerns and the open nature of the models can be challenging. At the application layer, developers may have more control and visibility over how their applications are being used, making it somewhat easier to monitor for misuses and unintended consequences. Nonetheless, the scale and complexity of monitoring efforts can still be resource-intensive and challenging to manage effectively.

### Implement Incident Reporting Channels

ACTORS: MODEL PROVIDERS, MODEL HOSTING SERVICES, APP DEVELOPERS
SECONDARY ACTOR: MODEL ADAPTERS

Actors from model providers,  to application developers, and other actors should implement secure channels for external stakeholders to report safety incidents or concerns. They should also enable internal teams to responsibly report incidents, potentially implementing whistleblower protection policies. Additionally, actors could contribute appropriate anonymized data to collaborative incident tracking initiatives like the AI Incident Database to enable identifying systemic issues, while weighing trade-offs like privacy, security, and other concerns. However, the effectiveness of incident reporting channels relies on stakeholders being aware of and willing to use them, which may require ongoing education and trust-building efforts.

## RESPOND

### Enforce Consequences for Policy Violations

ACTORS: MODEL HOSTING SERVICES, APP DEVELOPERS

Model hosting services and app developers should enforce consequences for users who violate their terms of use or engage in misuse of the hosted models. This can include issuing warnings, suspending or terminating access, requiring changes to AI projects, limiting discoverability from search engines or recommendation systems, and reporting severe cases to relevant authorities. A company's terms of use should clearly outline the acceptable use of its models and the consequences for violations. Detecting and enforcing consequences for acceptable use policy violations in open models may be more difficult for model hosting services due to the decentralized nature of access and use. Enforcement relies on user logins, and so more effectively governs registered users uploading models than it does others downloading models.

### Establish Decommissioning and Incident Response Policies

ACTORS: MODEL PROVIDERS, MODEL HOSTING SERVICES
SECONDARY ACTOR: MODEL ADAPTERS

Model providers and hosting services should establish decommissioning policies to recall a model, including criteria for determining when to stop hosting a model or when to adopt changes to the model's license to limit or prohibit continued use or development. They should consider when to responsibly retire support for foundation models based on well-defined criteria and processes. It's important to note that after open release of a foundation model's weights, its original developers will in effect be unable to decommission AI systems that others build using those model weights.

### Develop and Adhere to Transparency Reporting Standards

ACTORS: MODEL PROVIDERS, MODEL HOSTING SERVICES, APP DEVELOPERS

As commercial uses evolve, model providers, hosting services, and application developers could participate in collaborative initiatives with industry, civil society, and academia to align on transparency reporting standards for model usage. They could release periodic transparency reports following adopted standards, disclosing aggregated usage statistics and violation data while ensuring user privacy and data protection. These reports could provide insights into the scale and nature of misuse incidents and the actions companies taken to address them. For models that are downloaded and run locally, monitoring or reporting may be infeasible since users can run them on their own devices. However, the extent to which users prefer using cloud-based versions of models over running them locally, for example, due to the hardware and expertise required to run them, is unclear. In such cases, hosting services, rather than the open model providers, should consider releasing transparency reports. However, developing and adhering to transparency reporting standards may be especially challenging for open models given their decentralized nature.

## Endnotes

[1] "Risk Taxonomy." PAI Guidance for Safe Foundation Model Deployment. Partnership on AI, 2023.

[2] Henderson, Peter, et al. "Safety Risks from Customizing Foundation Models via Fine-Tuning." Policy Brief. Stanford Human-Centered Artificial Intelligence, January 11, 2024.

[3] Lakatos, Santiago. "A Revealing Picture: AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business." Graphika, December 2023.

[4] Widder, David Gray, and Dawn Nafus. "Dislocated accountabilities in the 'AI supply chain': Modularity and developers' notions of responsibility." Big Data & Society (January-June 2023).

[5] Cobbe, Jennifer, Michael Veale, and Jatinder Singh. "Understanding accountability in algorithmic supply chains." ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).

[6] "GitHub Response to NTIA Request for Comment on 'Dual-Use Foundation Artificial Intelligence Models With Widely Available Model Weights'." GitHub, March 27, 2024.

[7] "Policy Readout — Columbia Convening on Openness and AI." Mozilla Foundation, March 27, 2024.

[8] "New commitment to deepen work on severe AI risks concludes AI Seoul Summit." UK Department for Science, Innovation and Technology, May 22, 2024.

[9] Bommasani, Rishi, and Sayash Kapoor, et al. "Considerations for Governing Open Foundation Models." Issue Brief. Stanford Human-Centered Artificial Intelligence, December 2023