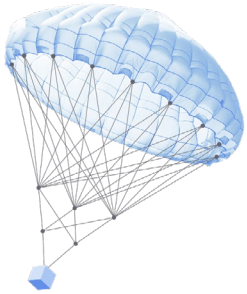
PARTNERSHIP ON AI



PAI's Guidance for Safe Foundation Model Deployment

MODEL TYPE
Advanced Narrow and General Purpose

RELEASE TYPE
Open Access

Partnership on AI's Guidance for Safe Foundation Model Deployment, first released in 2023, provides a framework for model providers to responsibly develop and deploy AI models. Following extensive stakeholder feedback during our public comment period highlighting the need to examine roles and responsibilities across the AI value chain, PAI has produced expanded guidance addressing key actors beyond model providers, including model adapters, model hosting services, and application developers, with particular focus on open foundation models.

Recent advances in foundation models have transformed the AI landscape, enabling content generation and paving the way for interactive systems that will be capable of performing complex digital tasks autonomously. While these models offer unprecedented opportunities for scientific discovery, productivity enhancement, and creative expression, they also present complex challenges including potential misuse, novel risks from increasingly capable systems, and the need for robust safety measures.

The Model Deployment Guidance website provides guidelines for various model capability and release type combinations. The Guidance also addresses significant model updates that expand capabilities post-deployment, requiring renewed governance processes.

The guidelines scale according to model capabilities and release types, with more extensive requirements for more capable models and widely available releases. This framework is meant to inform emerging regulatory frameworks, including the EU's general purpose AI Code of Practice, while providing practical guidance for safety measures companies should invest in developing.

Given the potentially far-reaching impacts of foundation models, translating shared safety principles into practical guidance requires collective action. These frameworks represent ongoing collaboration between industry, civil society, academia, and government to establish effective, collectively-agreed upon practices for responsible AI development and deployment.

This Guidance Checklist is one of three targeted frameworks that address distinct development and deployment scenarios.

Please see the other two frameworks:

Frontier x Restricted Release

For paradigm-shifting foundation models requiring extensive safety measures

Frontier x Closed Deployment

For internal deployments where models are directly integrated into products without public release

Definitions

FOUNDATION MODELS

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.

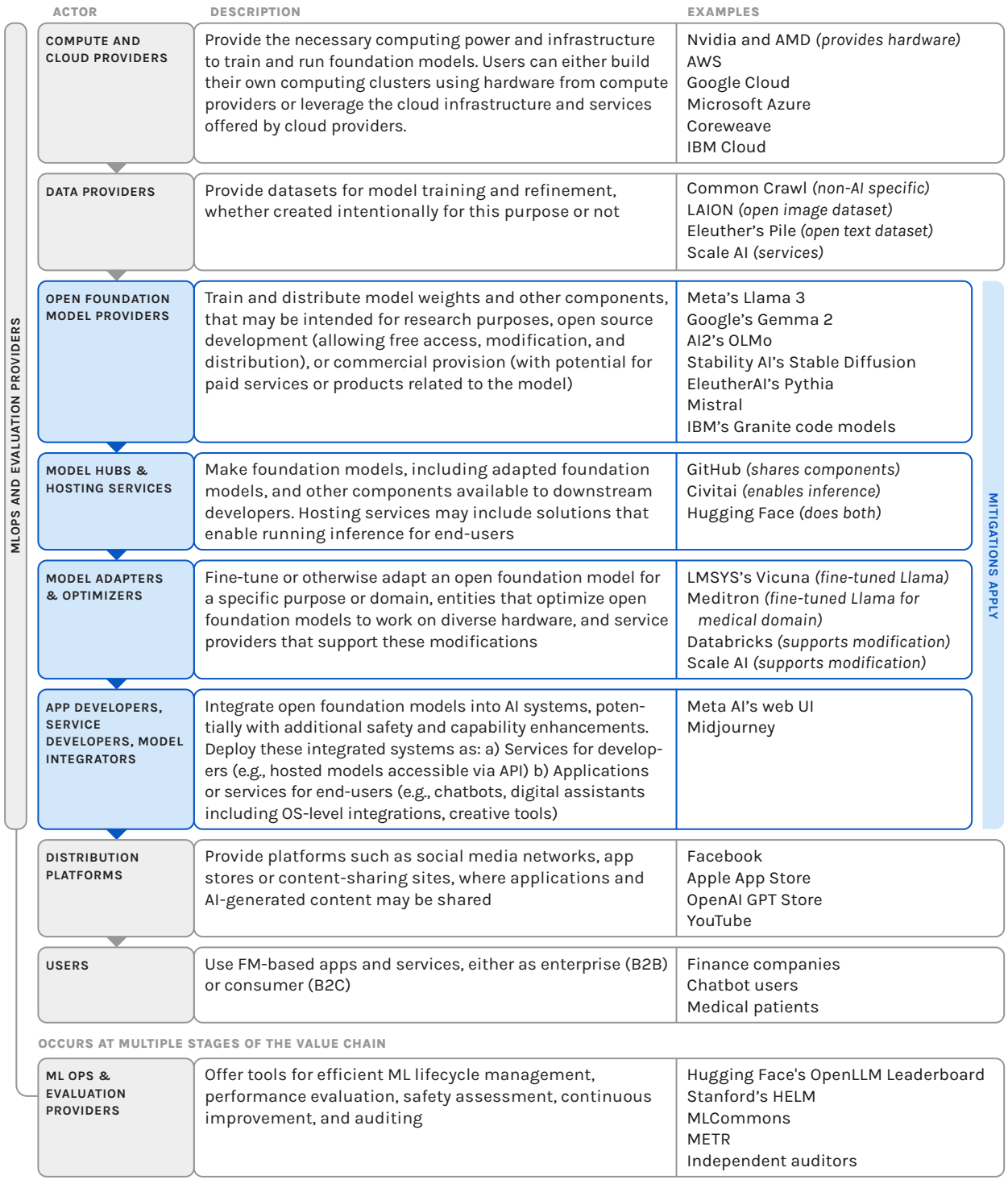
MODEL PROVIDERS

Model providers are organizations that train foundation models and distribute their components (such as model weights), that others may build on. These providers may operate with different objectives and distribution approaches:

- Research purposes (enabling scientific investigation and advancement)
- Open source development (allowing free access, modification, and distribution)
- Commercial provision (offering paid services or products)

The Open Foundation Model Value Chain

Given the decentralized development and deployment of open foundation models, risk mitigation responsibilities should be distributed across multiple actors. This guidance addresses key stakeholders including model providers who train the initial models, model adapters who fine-tune them, hosting services that make them available, and application developers who integrate them into products. The diagram below shows this complex web of actors involved in developing and deploying these models.



Guidance Checklist

MODEL TYPE

Advanced Narrow and General Purpose

Models with generative capabilities for synthetic content (text, image, audio, video).

Two subtypes:

1. Narrow purpose: Focused on specific tasks/modalities or focused on high-consequence domains (scientific, biological)
2. General purpose: Capable across diverse contexts (like chatbots/LLMs and multimodal models)

RELEASE TYPE

Open Access

Models released publicly with full access to key components, especially model weights. Can also include access to code, and data. Can be free or commercially licensed. Access can be downloadable or via cloud APIs and other hosted services.

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

- Proactive technical and policy measures 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.

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 ACTOR
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

ACTOR
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

ACTOR
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 ACTOR
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

- 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.

Monitor Misuses, Unintended Uses, and User Feedback

ACTORS
Model Providers
App Developers

SECONDARY ACTOR
Model Adapters

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 ACTORS
Model Adapters
Model Hosting Services

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

- 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.

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.