PARTNERSHIP ON AI

Documenting the Impacts of Foundation Models

A Progress Report on Post-Deployment Governance Practices

Jacob Pratt
Albert Tanjaya

Contents

Executive summary

Introduction

Foundation models are increasingly being deployed and adopted by society, but we still have limited data on their impact.

Web pages related to ChatGPT were viewed over three billion times in January 2025, making it the 13th most viewed domain on the internet, while billions of people use products deploying foundation models, with Gemini in Google Search and Meta AI in Meta products.[A] However, we are just beginning to understand the use and societal impact of these models after they have been deployed — i.e., post-deployment. What are these systems being used for? How are they helping people do things better? What are the most common or severe harms they cause?

There is a gap in research and policymaking[B] attention on the documentation of post-deployment impacts.

Policymaking and industry initiatives have focused on ensuring foundation models are "safe" to deploy, resulting in greater alignment in documentation and governance practices leading up to a model's release. However, society's understanding of these systems' post-release impacts is nascent. It is widely understood that the efficacy of our current focus on model evaluations and benchmarks is time-bound and insufficient, given that these approaches do not directly correlate to what makes these systems "safe" for deployment in complex societal systems. It is also difficult to establish clear frameworks for determining who should track and disclose specific post-deployment impacts in this complex system.

This post-deployment information gap prevents effective assessment of the benefits and risks of foundation models.

Collecting and sharing information on post-deployment impacts has faced significant challenges, including data-sharing and structural barriers, and a lack of norms and templates. What information should be collected, by whom, and how should it be documented and shared? Currently, policymakers, civil society, and academic groups have limited visibility into model usage patterns, impacts, serious incidents, and user feedback. This presents two challenges: policymakers face difficulties in evaluating risks and harms without reliable information, and organizations remain cautious about deploying models with uncertain trustworthiness. As highlighted in A Path for Science- and Evidence-based AI Policy, there is a need to "better understand AI risks" and "develop techniques and tools to actively monitor post-deployment AI harms and risks."

[A] Estimated traffic data for the chatgpt.com domain from similarweb.com of 3.7B site visits.

[B] This report uses "policy" to refer to voluntary or binding norms, rules, and frameworks that are developed by intergovernmental organizations, such as the UN and G7, and regional and national governmental organizations, such as the EU Commission and NIST, to achieve specific outcomes. "Policymakers" refers to the people and organizations that develop policies. A description of the AI governance stack with example policy initiatives and policymakers is in Appendix 2.

Goals for this report

Given the need for further work in understanding and assessing the post-deployment impact of foundation models, this work aims to achieve two core goals:

1. **Improve our collective understanding** of practices related to the documentation of post-deployment impacts.
2. **Drive accountability** for their adoption.

To achieve this, we explore the questions below and organize findings into the following sections:

Section 1	**Post-deployment documentation** • What practices contribute to the documentation of post-deployment impacts? • What are the benefits of these practices for different stakeholders?
Section 2	**Current state** • What early progress has the field made in adopting these practices?
Section 3	**Challenges** • What are the challenges of adopting these practices?
Section 4	**Recommendations** • Based on this understanding of the field, what can different stakeholders do to move the field forward? • What open questions remain to advance better practices?

An executive summary of these findings is presented below.

The current state of post-deployment impact documentation

Collecting, aggregating, and sharing post-deployment impact information publicly or privately provides four main benefits to actors in the foundation model value chain and other stakeholders:

1. **Amplifies societal benefits.** Documenting post-deployment impacts increases awareness of foundation model benefits and improves stakeholder literacy while building trust.
2. **Manages and mitigates risks.** Documenting post-deployment impacts enables stakeholders to identify, assess, and mitigate potential negative effects of AI systems on society.
3. **Develops evidence-based, proportionate policy.** Documenting post-deployment impacts provides policymakers with crucial data to develop and implement effective, balanced regulations and governance frameworks that protect people while considering implementation costs.
4. **Advances documentation standards through shared learning.** Multistakeholder collaboration in sharing post-deployment impact documentation helps establish best practices and moves the industry toward standardized approaches.

Building on academic literature, industry practices, and previous Partnership on AI (PAI) research, we hosted working group discussions and a workshop to identify what practices contribute to the documentation of post-deployment impacts and what their benefits are.

These multistakeholder activities highlighted **four key practices** with supporting processes that contribute to their adoption of post-deployment impact documentation. These processes do not fully operationalize a practice but were recognized as important during stakeholder discussions.

We then conducted a landscape review to identify where these practices were currently being adopted. While the AI value chain involves numerous actors, model providers are key decision-makers for a model's development, deployment, and distribution, and are well-positioned to collect, aggregate, and analyze information about their systems' impacts. Therefore, our primary focus is on how model providers can operationalize documentation practices.

The findings from this work are summarized below.

PRACTICE 1
Share usage information

Definition

Documenting information on how foundation models are used by downstream stakeholders. This practice might disclose the following information:[1]

- Activity data (input data; output data)
- Usage by geography
- Usage by sector, including high-risk sectors
- Usage by use case
- Total chat time usage
- Information on downstream applications

Processes

This practice involves the following processes:

- 1.1: Conduct surveys or user research to understand downstream usage.
- 1.2: Create tools to support the sharing of activity logs with trusted third parties for analysis.
- 1.3: Implement and track watermarking or identifiers.
- 1.4: Report aggregate usage statistics across geographies, sectors, or use cases, including usage in high-risk use cases.
- 1.5: Share information on downstream applications of the model.

Examples

- Anthropic's usage insights with Clio and Economic Index
- DSA transparency reports (e.g., Apple's)
- WildChat: ChatGPT Interaction Logs

Key findings

What is the current state of the field?

- There is low evidence of tracking and sharing of open model usage.
- There is low evidence of model usage being shared by restricted access model providers, though Anthropic's reporting on usage provides an early example to build on.
- External stakeholders are driving usage reporting through surveys, investigative reporting, and usage dataset creation.

Other considerations

- Actors can collect usage information by sharing data across the value chain, responsibly collating usage data, implementing and tracking identifiers, or conducting usage research (such as surveys).
- It's critical to account for differences between model release types, acknowledging that data collection mechanisms may vary.
- Application deployers may be well-positioned to collect usage information, while model providers may be better suited to aggregate and share usage information.

PRACTICE 2

Enable and share research on post-deployment societal impact indicators

Definition

Documenting and analyzing measurable indicators within the complex ecosystem where foundation models are deployed, recognizing that while direct attribution of impacts to specific models may not be possible, tracking key indicators can help understand emerging patterns and potential effects. This practice might disclose the following information:[2]

- **Labor Impact Indicators** (i.e., data-sourcing related risks and opportunities, task-related risks and opportunities, and workforce risks and opportunities). For more details, see PAI's Guidelines for AI and Shared Prosperity.
- **Environmental Impact Indicators** (i.e., compute, emissions, energy, and water usage from hardware and data centers, and geographical spread of data centers).
- **Synthetic Content Impact Indicators** (i.e., indirect/direct disclosure mechanisms, metrics related to the number of interactions with synthetic content labels, etc.). For more details, see PAI's Responsible Practices for Synthetic Media.

Processes

This practice involves the following processes:

- Aggregation of data usage for a specific model or specific field (See "Share Usage Information" Practice Section). *This is a prerequisite to enabling this practice.*
- 2.1-2.3: Reporting against labor, environmental, and synthetic content impact indicators.
- 2.4: Collaboration between actors across the value chain to enable model and data access for research purposes.
- 2.5: Dedication of resources and funding to aid research efforts in understanding societal, economic, or environmental impacts.

Examples

- Case studies of real-world examples operationalizing the Responsible Practices for Synthetic Media (Multi-industry case studies in collaboration with a civil society organization)
- AI brings soaring emissions for Google and Microsoft, a major contributor to climate change (NPR)

Key findings

What is the current state of the field?

- Impact-level stakeholders, such as civil society, academia, and other watchdog organizations, lead and contribute significantly to labor and environmental impact research.
- Some indicators, such as compute and data-sourcing information, can be found on established transparency artifacts like model cards.

Other considerations

- Model providers can support impact-level stakeholders in conducting societal impact research through various means, such as model access, data access, and researching funding opportunities.
- Different types of impacts will require different timelines for assessment, information, and data, and varying degrees of measurement.

PRACTICE 3
Report incidents and disclose policy violations

Definition

Documenting information on safety incidents and violations of policies and terms of use. This practice might disclose the following information:[3]

- Safety incidents (including records and summarized analysis)
- Violations of terms of use and policies (including records and summarized analysis)
- Mitigation and remediation actions

Processes

This practice involves the following processes:

- 3.1–3.2: Monitor for incidents and policy violations.
- 3.3: Share summaries of internal incident and policy violation reports.
- 3.4: Systematically report AI incidents to a third party (by actors across the value chain).[C]

Examples

- Google's voluntary and EU-mandated transparency reports
- AI Incident Database and incident summaries[D]
- OECD AI Incidents Monitor
- Common Vulnerabilities and Exploits Program (CVE)

Key findings

What is the current state of the field?

- Restricted access model providers conduct monitoring for policy violations, generally termed "abuse monitoring," and have uniform processes for reporting software-related security incidents, based on coordinated vulnerability disclosures.
- Monitoring conducted by open model providers will be different to monitoring conducted by restricted access model providers, and open model providers face unique considerations in conducting this monitoring.
- There is low evidence of organizations sharing summaries of findings from monitoring.
- User-reported, third-party AI incident databases exist, but there is limited coordinated reporting infrastructure.

[C] Using the OECD's definition around "AI incident" and "serious AI incident." A "serious" AI incident is related to the severity of the AI incident as defined by the OECD.

[D] PAI was the founding partner of the AI Incident Database.

PRACTICE 4
Share user feedback

Definition

Documenting and sharing feedback received on the model through a provider's feedback mechanism. This practice might disclose the following information:[4]

- Disclose the prompt given to a model and the response from the model specifically for problematic content related to criminal or regulated activity.
- Disclose the various types of feedback and ways to submit feedback based on the device, and how the provider uses the feedback received.

Processes

This practice involves the following processes:

- 4.1: Disclose the process for implementing a feedback mechanism for different stakeholders.
- 4.2: Aggregate individual feedback records to have as summaries.
- 4.3: Disclose the feedback follow-up process or, if warranted, the redress mechanism process.
- 4.4: Create incentive structures to invite stakeholders to participate in the feedback process proactively.

Examples

- Llama Output Feedback
- Send feedback with Gemini Apps

Key findings

What is the current state of the field?

- There are feedback mechanisms for users across model providers, such as providing email addresses, feedback forms, and community portals.
- Feedback collections can be seen on community or developer portals, but proactive summaries from model providers are not typically seen.
- There are few incentive structures for providing feedback, though bug bounties act as incentives for highlighting specific issues. It is possible that there are other incentive structures, however, they are not disclosed publicly.
- There is low disclosure of response or redress mechanisms after providing feedback. At the moment, it is possible that there are no disclosure mechanisms that exist, or the mechanisms that exist are not disclosed.

Across these four practices, model providers collect and potentially assess some information on post-deployment impacts. However, model providers generally do not share aggregated or granular information on these impacts.

Challenges

We also explored barriers hindering the adoption of these practices and specific processes. We found that the main challenges for collecting and sharing information on post-deployment impacts can be grouped into five themes:

CHALLENGE 1
Lack of standardization and established norms

The absence of agreed standards, templates, and responsibilities for documenting post-deployment impact information across industry and state actors creates uncertainty about who should implement these practices and how they should do it.

This often leads to "finger-pointing" — without a common agreement on responsibilities and accountability structures to promote the adoption of practices, it's easy to point the blame and responsibility to other organizations.

CHALLENGE 2
Data sharing and coordination barriers

Model providers create foundation models but may lack access to deployment data; app developers generate data but may be unable to share it due to privacy restrictions and competitive concerns; users generate impact data but have privacy protections; and researchers and policymakers need comprehensive data but may be unable to access it.

Different stakeholders have varying levels of access to post-deployment data, and the complications around privacy concerns and contractual barriers present operational challenges around coordination and data sharing.

CHALLENGE 3
Misaligned incentives

There are tensions between the drive for post-deployment transparency, protection of trade secrets, and potential legal or market scrutiny that could arise from disclosing post-deployment impacts.

CHALLENGE 4
Limited Infrastructure

Structural initiatives to coordinate information on post-deployment impacts, such as implementing reporting structures and establishing databases for analysis, have yet to be put in place.

CHALLENGE 5
Decentralized nature of open model deployment

The decentralized nature of open models creates unique considerations for model providers. The variety of stakeholders who may use, adapt, and build on open models, of various levels of "openness," creates barriers to collecting and sharing information on post-deployment impacts. This means different methods may be required to collect relevant information on post-deployment impacts.

Recommendations

To overcome these challenges and ensure the benefits of effective post-deployment governance, we recommend that stakeholders take the following actions to move the field forward.

GENERAL RECOMMENDATION

Implement processes that support sharing usage and feedback information, conducting societal impact research, and reporting incidents and policy violations

Organizations can improve transparency by documenting some post-deployment impacts now. These activities have examples that can be built on, or would benefit from initial testing of approaches. There is also a need for broader sociotechnical research into these models' societal impacts, supported by sustained commitments from model providers, governments, and academic institutions.

Key actions

- Model-level and application-level stakeholders should publish available information on violations, incidents, usage, activity logs, feedback, and environmental impacts, in an aggregated format, publicly and with trusted third parties.
- Organizations and governments should fund and conduct societal impact research, including independent research.

RECOMMENDATION 1

Define norms for the documentation of post-deployment impacts through multistakeholder processes, which may be formalized through technical standards

Addresses Challenge 1: Lack of norms and responsibilities

A multistakeholder approach is needed to develop shared norms and standards for post-deployment impact documentation, though some areas still require foundational research before standardization can occur. Existing research, unofficial standards, and organizations like ISO/IEC and NIST can support this process by ensuring interoperability across different frameworks and policies.

Key actions

- Stakeholders should conduct research into methods for documenting post-deployment impacts.
- Stakeholders should contribute to multi-stakeholder standards development processes.
- Stakeholders should encourage interoperability between standards and policies.

RECOMMENDATION 2
Explore mechanisms for responsible data sharing

Addresses Challenge 2: Data sharing and coordination barriers

Responsible data sharing of post-deployment impact information requires privacy-preserving mechanisms, an environment conducive to data sharing, and specific procedures for academic access. Model providers and other actors should establish contractual agreements for data sharing and create equitable access for academic researchers and civil society, potentially through legal safe harbors and tools that enable user consent for research datasets.

Key actions

- Model- and application-level stakeholders should explore how technical privacy-preserving mechanisms can be applied to sharing documentation on post-deployment impacts.
- Model- and application-level stakeholders should foster environments for sharing information for disclosure.
- Model- and application-level stakeholders should explore options for academic and civil society data access procedures.

RECOMMENDATION 3
Policymakers should explore where guidance and rules on documenting post-deployment impacts are needed

Addresses Challenge 3: Misaligned incentives

Policymakers play a crucial role in mandating the documentation of post-deployment impacts where industry incentives conflict with societal interests and should assess where policies are needed to counterbalance these incentives. They should also explore how this information can inform policymaking and encourage interoperability where possible.

Key actions

- Policymakers should review how documentation on post-deployment impacts can help them identify and assess risks.
- Policymakers should identify where current adoption levels of post-deployment impact documentation conflict with policy objectives.
- Policymakers should explore how guidance and rules for monitoring and documenting post-deployment impacts can align with the EU Code of Practice.

RECOMMENDATION 4
Policymakers should develop blueprints for national post-deployment monitoring functions

Addresses Challenge 4: Limited infrastructure

Well-resourced, legitimate structures for collating and analyzing post-deployment impact information will help stakeholders identify and assess systemic risks. Policymakers are well-positioned to develop and legitimize these structures.

Key actions

- Policymakers should commit to investing in the capacity, capabilities, and structures that will enable post-deployment monitoring of foundation models by September 2025.
- Policymakers should develop blueprints for post-deployment incident monitoring of foundation models by March 2026.
- Policymakers should build on incident reporting structures to monitor the usage of foundation models that exceed agreed-upon risk or capability thresholds by March 2027.

RECOMMENDATION 5

Conduct research on methods for collecting information about open model impacts

Addresses Challenge 5: Decentralized nature of open model deployment

Open model providers should research how to responsibly document post-deployment impacts and adopt a multistakeholder approach by collaborating with academia and civil society organizations. This research should address challenges around user privacy, security, and trust-building while establishing clear responsibilities among the many stakeholders involved with open models.

Key actions

- Open model providers should collaborate with academia, civil society, and other open model providers to research the methods for collecting usage information and identifying and sharing societal impact indicators.
- Open model providers should collaborate with application developers and model-hosting services to determine realistic responsibilities for monitoring incidents and policy violations for open models.

Key questions

Beyond these recommendations, there are key questions that we encourage further research and discussion into:

- For which issues, whether due to urgency or public interest, should policymakers lead the definition of best practices and develop binding rules?
- How can specific processes be implemented for open models?
- What additional practices for documenting post-deployment impacts are important to foundation model governance?
- What level of detail is required in documentation to measure relevant impacts and assess trustworthiness?
- What counts as a substantial modification to a model? When an organization substantially modifies a model, what are its responsibilities for documenting post-deployment impacts, and how can this administrative burden be managed?

Looking ahead

This report will inform the topics that PAI will explore in 2025, and how PAI updates tools, methods, and guidance related to documentation, transparency, and safety. PAI will also continue to improve collective understanding of the field and drive accountability through future progress reports. If you want to know more, please contact PAI's Public Policy team at policy@partnershiponai.org.

Introduction

Documentation and transparency play a key role in ensuring the responsible development and deployment of foundation models, as noted in PAI's recent Policy Alignment on AI Transparency report. This recognition is highlighted in international principles by the OECD, regional legislation such as the EU AI Act, industry standards such as Microsoft's, and academic literature.

In the last 18 months, policymakers[A] at the international and regional levels have shifted their attention to foundation models, with documentation and disclosure requirements as core focuses of their efforts.[B] As these policy and regulatory initiatives go into effect and continue to be built out, they will give model providers increasingly strong incentives to adopt specified documentation templates across the AI model's development lifecycle. One example is the training data template to be developed through the EU's Code of Practice process, which will boost transparency in areas of regulatory attention. Much of the current regulatory focus has been on research and development and pre-deployment phases with common requirements, including the disclosure of risk management processes and assessments, training processes, model information, evaluations, capabilities, and limitations. This ex-ante approach is directed at ensuring that future models are designed and developed responsibly, with the appropriate protections in place.

As global policy frameworks continue to take shape, model providers can build upon established documentation practices while remaining adaptable to new requirements and standards. Information on a model's modality, training data, red teaming results, usage policies, safety safeguards, and many other common requirements mentioned above can often be found in accepted industry artifacts such as model cards, factsheets, and system cards. However, these current practices and artifacts do not always identify or assess the real-world impacts models have post-deployment. There is currently a lack of agreed standards and other challenges, detailed in Section 3: Challenges, specifying who documents, what to document, and how to document and share information about a model's impact post-deployment.

[A] This report uses "policy" to refer to voluntary or binding norms, rules, and frameworks that are developed by intergovernmental organizations, such as the UN and G7, and regional and national governmental organizations, such as the EU Commission and NIST, to achieve specific outcomes. "Policymakers" refers to the people and organizations that develop policies. A description of the AI governance stack with example policy initiatives and policymakers is in the Appendix.

[B] See Table 2B "Comparison of documentation requirements across in-scope frameworks" of PAI's Policy Alignment on AI Transparency.

Goals for this report

Given the need for further work in understanding and assessing the post-deployment impact of foundation models, this work aims to achieve two core goals:

1. **Improve our collective understanding** of practices related to the documentation of post-deployment impacts.
2. **Drive accountability** for their adoption.

This report builds on previous PAI guidelines to achieve these two goals. PAI has long standing work in documentation and post-deployment governance practices, with PAI's ABOUT ML work providing early guidance for documentation and PAI acting as the founding partner of the AI Incident Database. Post-deployment practices were recommended in PAI's Guidance for Safe Foundation Model Deployment. This work primarily builds on the "develop transparency reporting standards" and relevant societal impact guidelines, exploring several baseline and recommended practices that could support the development and sharing of information about post-deployment impacts. These practices include:

- Release periodic transparency reports following established standards, disclosing aggregated usage insights and violation data. Take appropriate measures to ensure transparency reporting protects user privacy and data.
- Assess downstream real-world impact of models, for example, in collaboration with external researchers.
- Contribute appropriate anonymized data to collaborative incident tracking initiatives to enable identifying systemic issues, while weighing trade-offs like privacy, security, and other concerns.
- Monitor and report on environmental impacts of model development and deployment.
- Provide transparency into monitoring practices while protecting user privacy.
- Collaborate across industry, civil society, academia, and worker organizations to advance the measurement, responsible disclosure practices, and mitigation of severe labor market risks.

This work also builds on findings from PAI's recent Policy Alignment on AI Transparency report, which highlighted calls for post-deployment monitoring and documentation artifacts. There has been high-level policy recognition of the importance of the practices detailed above, but there has been little action to drive the adoption of these practices forward. This report represents a series of first steps towards driving adoption by exploring how progress can be made through the implementation of these practices.

To achieve the goals, we explore the following questions and organize findings into four sections:

Section 1	**Post-deployment documentation** • What practices contribute to the documentation of post-deployment impacts? • What are the benefits of these practices for different stakeholders?
Section 2	**Current state** • What early progress has the field made in adopting these practices?
Section 3	**Challenges** • What are the challenges of adopting these practices?
Section 4	**Recommendations** • Based on this understanding of the field, what can different stakeholders do to move the field forward? • What open questions remain to advance better practices?

Documenting post-deployment impacts is a responsibility shared by actors across the AI value chain. However, model providers make key decisions about a model's development, deployment, and distribution and are well-placed to collect, aggregate, and analyze information about the impact of their systems. Therefore, our primary focus is on how model providers can operationalize documentation practices, and we assess the field through this lens. Other actors may adopt documentation practices from this report as relevant.

How to read this report

This report is designed to be useful to different actors with different goals.

IF YOU ARE…	YOU CAN USE THIS REPORT...
a policymaker	as an evidence base to inform policy development related to documenting post-deployment impacts
	to identify key actions to explore over the next 6–24 months.
a model provider, model deployer, or actor in the value chain	to understand how documenting post-deployment impacts contributes to responsible model development and deployment
	to identify good practices from other providers that merit further exploration
an academic or civil society researcher or interested user of AI models and systems	to understand the field of documentation of post-deployment impacts for foundation model actors
	to identify open questions that merit further exploration

Looking ahead

Tracking the implementation of governance practices is as crucial as exploring what should be done — highlighting what practices are and are not being adopted ensures that stakeholders in the field are held accountable, and promotes positive use cases that can guide others. This aligns with PAI's core value of "transparency and accountability," supporting our mission to advance positive outcomes for people and society.

This report uses this starting point to explore the challenges in documenting post-deployment impacts and charts a route forward for the field. While organizations have made good progress in documenting model-related, evaluation-related, and other types of information, we identified a critical gap in the documentation of post-deployment impacts, as first described in PAI's Guidance for Safe Foundation Model Deployment. We focused on identifying specific adoption gaps and highlighting policy, technical, and other actionable solutions to improve adoption in this area, recognizing its importance as models continue to be deployed and policies evolve.

Findings from this report will inform the topics that PAI will explore in 2025, and how PAI updates tools, methods, and guidance related to documentation, transparency, and safety.

In consultation with PAI's Policy Steering Committee of global experts, PAI will also review how progress reports can continue to contribute to accountability by describing the state of the field, identifying gaps, and highlighting where action is needed. This will become more important as foundation models are more widely deployed, governance practices mature, and international policy develops in 2025.

Acknowledgements

PAI's work continues to be developed through multistakeholder consultation, and we are pleased to have delivered this work with the support of the PAI Policy Steering Committee, PAI's working group on the documentation of post-deployment impacts, and many others (see Appendix 3 for names) through six working group sessions, one workshop, and supporting discussions.

SECTION 1

Documentation on post-deployment impacts and why it matters

Who contributes to collecting post-deployment information and driving accountability?

Collecting information on post-deployment impacts requires collaboration across the value chain, and how this can be done will vary depending on the model's release strategy (such as an open or restricted model release). There are also other roles involved in the deployment of applications or services using these models, which can result in different deployment configurations and collaboration requirements for documenting post-deployment impacts.

TABLE 1. **Deployment configurations with example actors and products**

ACTORS	MODEL & SERVICE PROVIDER (OPENAI PROVIDES CHATGPT)	RESTRICTED ACCESS MODEL (OPENAI HOSTS GPT-4 API)	OPEN MODEL (AZURE HOSTS MISTRAL LARGE)
COMPUTE, CLOUD, & DATA PROVIDERS	Hardware and data provider	Hardware and data provider	Hardware and data provider
MODEL PROVIDERS	OpenAI	OpenAI	Mistral (Large)
MODEL HUBS & HOSTING SERVICES			Microsoft Azure
APP DEVELOPERS, SERVICE DEVELOPERS, MODEL INTEGRATORS		MyThorch	CodeSage
USERS	User	User	User

Adapted from The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI

This means understanding the value chain and wider accountability ecosystem is important for feasibly documenting post-deployment impacts. Building on the stakeholders identified in PAI's Risk Mitigation Strategies for the Open Foundation Model Value Chain, we highlight four main groups of stakeholders and provide more information in Appendix 1.

TABLE 2. **Four categorizations of stakeholders from the model development value chain**

STAKEHOLDER CATEGORY	RELEVANCE IN THE POST-DEPLOYMENT SETTING	STAKEHOLDER TYPE
1. Hardware and Data Level Actors that provide the infrastructure for foundation model development such as compute hardware, cloud set-up, and dataset creators	They hold insights and data concerning factors impacting the environment, such as energy consumption, compute power, and usage, etc.	Compute & Cloud Providers Data Providers
2. Model Level Actors that directly develop, train, fine-tune, or optimize foundation models	They have the power to alter, remove, and add foundation models into the space, thus impacting society overall	Model Providers Model Hubs & Hosting Services Model Adapters & Optimizers
3. Application Level Actors that interact with foundation models (either the base models or the fine-tuned variations) to integrate into AI systems, conduct research, or host for distribution and access purposes	They help proliferate models by developing systems and services for widespread use or distributing them to a broader range of impact-level stakeholders	App Developers, Service Developers, Model Integrators Distribution Platforms Assurance Providers
4. Impact Level Actors affected by the use of AI models and systems in society who have some degree of external power to drive accountability for transparency and safety	It encompasses various actors affected by the deployment of foundation models	Users Civil Society & Other Watchdog Organizations Academia Regulators & Norm Enforcers AI Subjects/Data Rights Holders

It is important to note that:

- **A given actor will have differing degrees of power and responsibilities depending on a model's release type.**
  For example, when OpenAI deploys GPT-4 through their API, they have more access to understanding that model's usage downstream. They can also react immediately to any reports or incidents and issue changes if they arise. On the other hand, when Meta releases an open source model like Llama 3 through Hugging Face, their access to any downstream metrics is limited, shifting more responsibilities to application developers and model hosting services.
- **One actor may play the role of multiple actors.**
  For example, foundation model providers may also host, adapt, and build applications with their own models.

Why documentation on post-deployment impacts matters

Documentation and disclosures are long standing, necessary practices. Taking time to collect, assess, and document information about any part of a model's development, deployment, and impact supports responsible model development and use, catalyzes best practices, and bolsters ethical sensitivity and deliberation within model providers.[5] [6]

These benefits stem from the act of documenting information, and there are additional benefits from the subsequent use of that information by different actors in the ecosystem. In particular, sharing relevant information about a model's impact after deployment is crucial to understanding how to amplify societal benefits, manage and mitigate risks, develop evidence-based proportionate policy, and advance industry-wide norms.

Amplify societal benefit

Sharing post-deployment information can play a role in increasing awareness of the benefits of foundation model use. For example, relevant information may aid in identifying valuable use cases, supporting scale-up of use, as well as improving the literacy of stakeholders inside and outside of the value chain. It may also increase trust in these systems, encouraging responsible use.

Sharing information about a model's impact after deployment is crucial to understanding how to amplify societal benefits, manage and mitigate risks, develop evidence-based proportionate policy, and advance industry-wide norms.

Manage and mitigate risks

"Risks" are the potential negative impacts that AI models and systems have on society, and "harms" are actual negative impacts. To properly manage risks and harms, stakeholders need to identify a risk, assess the likelihood and severity of the risk, and mitigate that risk. Each stage requires data and evidence on the impact of models, which documentation for post-deployment impacts can provide.

We highlight risks that are common across risk frameworks in Appendix 2 and note that terminology can differ across domains, with "AI hazards" and "AI incidents" used by the OECD for incident monitoring.

Develop evidence-based, proportionate policy

Policymakers can deploy various regulatory tools to support their policy objectives related to protecting and benefiting people, businesses and the environment. However, developing, implementing, and enforcing regulation has a cost, so any intervention should be proportionate to the issue at hand. This assessment and decision-making requires data and evidence to be effective, illustrated by recent calls for evidence-based policymaking.[7]

Documenting post-deployment impacts provides the means to assess and identify benefits and risks more effectively. For example, analyzing usage data may help stakeholders assess the likelihood of the "malicious uses" risk by identifying the percentage or total number of prompts that result in non-consensual intimate imagery being generated. This can then support policymakers in developing rules that are proportionate to the issue.

Advance documentation standards through shared learning

Given the nascent state of post-deployment impact documentation, sharing insights and artifacts across stakeholders — even if their format and content vary — creates a valuable foundation for developing shared practices. Through multistakeholder forums, the industry can collectively learn from diverse experiences and identify common patterns in documentation needs, gradually moving toward more standardized approaches that serve all stakeholders effectively.

Documentation on post-deployment impacts

Though there are not yet agreed standards and frameworks for sharing information about post-deployment impacts, there is a base of academic literature that describes information that can be publicly or privately shared and the need for this information. To collect, collate, and share this information, model providers may need to adopt various processes in collaboration with other actors.

For each practice, the following is provided:

- **Summary:** A summary of what the practice entails.
- **Information:** Information that can be publicly or privately shared as part of this practice.
- **Benefits:** A summary of the benefits of adopting this practice.
- **Processes:** Activities that model providers, in collaboration with other actors, can undertake to operationalize this practice.
- **Examples:** Examples of this practice related to the deployment of foundation models or from other domains.

The processes were identified through a landscape review and workshop discussions and are not a complete set of activities to operationalize a practice. Challenges to fully adopting these processes are explored in Section 3: Challenges.

PRACTICE 1
Share usage information

Definition	Documenting information on how foundation models are used by downstream stakeholders. This practice might disclose the following information:[8] • Activity data (input data; output data) • Usage by geography • Usage by sector, including high-risk sectors • Usage by use case • Total chat time usage • Information on downstream applications
Processes	This practice involves the following processes: • 1.1: Conduct surveys or user research to understand downstream usage. • 1.2: Create tools to support the sharing of activity logs with trusted third parties for analysis. • 1.3: Implement and track watermarking or identifiers. • 1.4: Report aggregate usage statistics across geographies, sectors, or use cases, including usage in high-risk use cases. • 1.5: Share information on downstream applications of the model.
Examples	• Anthropic's usage insights with Clio and Economic Index • DSA transparency reports (e.g., Apple's) • WildChat: ChatGPT Interaction Logs

PRACTICE 2
Enable and share research on post-deployment societal impact indicators

Definition

Documenting and analyzing measurable indicators within the complex ecosystem where foundation models are deployed, recognizing that while direct attribution of impacts to specific models may not be possible, tracking key indicators can help understand emerging patterns and potential effects. This practice might disclose the following information:[9]

- **Labor Impact Indicators** (e.g., data-sourcing related risks and opportunities, task-related risks and opportunities, and workforce risks and opportunities). For more details, see PAI's Guidelines for AI and Shared Prosperity.
- **Environmental Impact Indicators** (e.g., compute, emissions, energy, and water usage from hardware and data centers, and geographical spread of data centers).
- **Synthetic Content Impact Indicators** (e.g., indirect/direct disclosure mechanisms, metrics related to the number of interactions with synthetic content labels, etc.). For more details, see PAI's Responsible Practices for Synthetic Media.

Processes

This practice involves the following processes:

- Aggregation of data usage for a specific model or specific field (See "Share Usage Information" Practice Section). *This is a prerequisite to enabling this practice.*
- 2.1-2.3: Reporting against labor, environmental, and synthetic content impact indicators.
- 2.4: Collaboration between actors across the value chain to enable model and data access for research purposes.
- 2.5: Dedication of resources and funding to aid research efforts in understanding societal, economic, or environmental impacts.

Examples

- Case studies of real-world examples operationalizing the Responsible Practices for Synthetic Media (Multi-industry case studies in collaboration with a civil society organization)
- AI brings soaring emissions for Google and Microsoft, a major contributor to climate change (NPR)

PRACTICE 3
Report incidents and disclose policy violations

Definition

Documenting information on safety incidents and violations of policies and terms of use. This practice might disclose the following information:[10]

- Safety incidents (including records and summarized analysis)
- Violations of terms of use and policies (including records and summarized analysis)
- Mitigation and remediation actions

Processes

This practice involves the following processes:

- 3.1–3.2: Monitor for incidents and policy violations.
- 3.3: Share summaries of internal incident and policy violation reports.
- 3.4: Systematically report AI incidents to a third party (by actors across the value chain).[A]

[A] Using the OECD's definition around "AI incident" and "serious AI incident." A "serious" AI incident is related to the severity of the AI incident as defined by the OECD.

Examples

- Google's voluntary and EU-mandated transparency reports
- AI Incident Database and incident summaries[B]
- OECD AI Incidents Monitor
- Common Vulnerabilities and Exploits Program

[B] PAI was the founding partner of the AI Incident Database.

PRACTICE 4

Share user feedback

Definition

Documenting and sharing feedback received on the model through a provider's feedback mechanism. This practice might disclose the following information:[11]

- Disclose the prompt given to a model and the response from the model specifically for problematic content related to criminal or regulated activity.
- Disclose the various types of feedback and ways to submit feedback based on the device, and how the provider uses the feedback received.

Processes

This practice involves the following processes:

- 4.1: Disclose the process for implementing a feedback mechanism for different stakeholders.
- 4.2: Aggregate individual feedback records to have as summaries.
- 4.3: Disclose the feedback follow-up process or, if warranted, the redress mechanism process.
- 4.4: Create incentive structures to invite stakeholders to participate in the feedback process proactively.

Examples

- Llama Output Feedback
- Send feedback with Gemini Apps

This list is not intended to be a comprehensive list of post-deployment practices, and some practices — such as sharing updates and decommissioning information — are not included. However, this selection includes key practices that provide significant societal benefits while also providing a manageable scope for the working group, workshop attendees, and PAI team to explore in detail.

How these practices should be adopted, particularly for open models, is an evolving area of discussion. For example, the second draft of the EU Code of Practice emphasizes the need to collect relevant post-deployment information and monitor real-world usage, but acknowledges that how to effectively monitor open models without negatively impacting downstream users is an "open question." This is discussed further in Section 3: Challenges.

What is the current state of the field?

To realize the benefits of documenting post-deployment impacts, we need to understand the current state of the field and identify areas where progress can be made. While the AI value chain involves numerous actors, model providers make key decisions about a model's development, deployment, and distribution, and are well-placed to collect, aggregate, and analyze information about the impact of their systems. Therefore, our primary focus is on how model providers can operationalize documentation practices.

Assessing the field requires a recognition of some key considerations[A] that affect what documentation processes should be adopted, how they should be implemented, and by whom:

- **The model release type affects how information can be collected and which actors should be involved.** See Table 1 and Challenge 5 for more discussion.
- **Not all information should be disclosed immediately and publicly.** There is significant benefit to sharing information openly, but some information should only be disclosed to trusted actors, for privacy, security, or other reasons. For example, publicly sharing a full incident report on a user data vulnerability may violate user privacy rights, and a carefully redacted disclosure may be more appropriate.

To assess the current state of the field in adopting documentation practices for post-deployment impacts, we looked for best practices by organizations providing the following models:[B]

OPEN MODELS	RESTRICTED ACCESS MODELS
Allen Institute for AI: OLMo	**Anthropic:** Claude 3.5
BigCode: StarCoder	**Cohere:** Command
IBM: Granite	**Google:** Gemini 1.5
Meta: Llama 3.1	**Inflection:** Inflection 2.5
Mistral: Mistral Small	**OpenAI:** GPT-4
Microsoft: Phi-3	
Stability AI: Stable Diffusion 3	

We describe the "level of adoption" of processes where model providers are key actors and highlight supporting actors. To assess the "level of adoption" of a process, we use the following guide:

[A] Additional considerations include:

- The model's capability will affect the information that should be disclosed. *(How "capability" should be assessed, and what thresholds should apply to distinguish requirements by "capability," is still a matter of debate in the field.)*
- Information shared privately or publicly should focus on audience needs and present this information accordingly.
- Disclosure of this information may inform changes in practice.
- Foundation models will be fine-tuned by downstream actors, and it can be difficult to define where responsibility for negative impacts lies.
- Some impact information is difficult to gather on a short- to medium-term timeline.

See Appendix 1 for more information.

[B] We reviewed models created in the US and EU, though their reach and impact affect a global audience.

LEVEL OF ADOPTION[C] (NOV. 2024)	PRACTICES IDENTIFIED (OUT OF 12)	INTERPRETATION
NONE	0	No organizations have implemented this process. Significant work is required to overcome the challenges blocking adoption.
LOW	1 — 3 *(or >0 partial implementations)*	Few organizations (≤25%) have implemented this process, or some organizations have partially implemented it. The field can build on these initial practices, but more work is needed.
MEDIUM	4 — 6	Some organizations (26% - 50%) have implemented this process. The field can learn from the good practices highlighted.
HIGH	7 — 12	Most organizations (>50%) have implemented this process. Future work should focus on identifying and aligning around best practices.

For one process, model providers may not play a key role in the process, so we do not measure the level of adoption and highlight this using ***N/A****.*

Supporting evidence and analysis is shared in this Google Sheets document.

A lower level of adoption may indicate a lack of feasibility or a lack of will to implement these processes. Analysis of which challenges are present for which processes is provided in Table 3, Section 3.

[C] We do not assess the quality of the adopted practices when describing the "level of adoption," but do explore them in more detail in the Annex. Limitations of this method are described in the "Methodology" section.

PRACTICE 1

Share usage information

Sharing information on how foundation models are used by downstream stakeholders

PROCESSES	LEVEL OF ADOPTION (NOV. 2024)	OPEN MODEL CHALLENGES	REQUIRES MODEL-LEVEL COLLABORATION	OTHER RELEVANT STAKEHOLDERS
1.1 Conduct surveys or user research to understand downstream usage	N/A		?	CS · A · R
1.2 Create tools to support the sharing of activity logs with trusted third parties for analysis	NONE	●	?	CS · A
1.3 Implement and track watermarking or identifiers	NONE			
1.4 Report aggregate usage statistics, across geography, sector or use case, including usage in high-risk use cases	LOW	●	●	
1.5 Share information on downstream applications of the model	LOW	●	●	

KEY

?	Warrants further exploration
CCD	Compute, Cloud or Data Providers
CS	Civil Society and Watchdogs
A	Academia
R	Regulators and Norm Enforcers

KEY FINDINGS

- Actors can collect usage information by sharing data across the value chain, responsibly collating usage data, implementing, and tracking identifiers, or conducting usage research (such as surveys).
- It's critical to account for differences between model release types, acknowledging that data collection mechanisms may vary.
- There is low evidence of tracking and sharing of open model usage.
- There is low evidence of model usage being shared by restricted access model providers, though Anthropic's reporting on usage provides an early example to build on.
- Application developers may be well-positioned to collect usage information, while model providers may be better suited to aggregate and share usage information.
- External stakeholders are driving usage reporting through surveys, investigative reporting, and usage dataset creation.

PRACTICE 2

Enable and share research on post-deployment societal impact indicators

Sharing and analyzing measurable indicators within the complex ecosystem where foundation models are deployed, recognizing that while direct attribution of impacts to specific models may not be possible, tracking key indicators can help understand emerging patterns and potential effects

PROCESSES	LEVEL OF ADOPTION (NOV. 2024)	OPEN MODEL CHALLENGES	REQUIRES MODEL-LEVEL COLLABORATION	OTHER RELEVANT STAKEHOLDERS
2.1 Report on labor impact indicators[D] *(i.e., data-sourcing related risks and opportunities, task-related risks and opportunities, workforce risks and opportunities)*	LOW	●	●	CS · A · R
2.2 Report on environmental impact indicators *(i.e., compute, emissions, energy, and water usage from hardware and data centers, and geographical spread of data centers)*	LOW [E]	●	●	CCD · CS · A · R
2.3 Report on synthetic content impact indicators[F] *(e.g., indirect/direct disclosure mechanisms, metrics related to the number of interactions with synthetic content labels, etc.)*	LOW [G]	●	●	CS · A
2.4 Disclosure of third-party research access	MEDIUM	●	●	
2.5 Disclosure of organizational resourcing commitments and dedicated funding commitments towards post-deployment societal impacts	HIGH		●	

KEY

CCD	Compute, Cloud or Data Providers
CS	Civil Society and Watchdogs
A	Academia
R	Regulators and Norm Enforcers

KEY FINDINGS

- Impact-level stakeholders, such as civil society, academia, and other watchdog organizations, lead and contribute significantly to labor and environmental impact research.
- Model providers are in the position to support impact-level stakeholders in conducting societal impact research through various means such as model access, data access, and research funding opportunities.
- Some indicators, such as compute and data-sourcing information, can be found on established transparency artifacts like model cards.
- Different types of impacts will require different timelines for assessment, information and data, and varying degrees of measurement.

[D] For more details, see PAI's Guidelines for AI and Shared Prosperity and system cards.

[E] Compute information for a majority of models is generally available on established transparency artifacts like model cards and system cards.

[F] For more details, see PAI's Responsible Practices for Synthetic Media.

[G] Low indicator here may dictate that not all models have the capacity to produce synthetic content.

PRACTICE 3

Report incidents and disclose policy violations

Sharing information on safety incidents and violations of policies and terms of use

PROCESSES	LEVEL OF ADOPTION (NOV. 2024)	OPEN MODEL CHALLENGES	REQUIRES MODEL-LEVEL COLLABORATION	OTHER RELEVANT STAKEHOLDERS
3.1 Monitor for incidents	MEDIUM [H]	●	●	
3.2 Monitor for policy violations	MEDIUM [H]	●	●	
3.3 Share summaries of internal incident and policy violation reports	LOW	●	●	
3.4 Systematically report AI incidents to a third party with respect to their severity	NONE		●	CS · A · R

KEY	
CCD	Compute, Cloud or Data Providers
CS	Civil Society and Watchdogs
A	Academia
R	Regulators and Norm Enforcers

H Adoption was assessed across these two processes, as "abuse monitoring" could cover both.

KEY FINDINGS

- Restricted access model providers conduct monitoring for policy violations, generally termed "abuse monitoring," and have uniform processes for reporting software-related security incidents, building on coordinated vulnerability disclosures.
- Monitoring conducted by open model providers will be different to monitoring conducted by restricted access model providers, and open model providers face unique considerations in conducting this monitoring.
- There is low evidence of organizations sharing summaries of findings from monitoring.
- Third-party AI incident databases exist for users to voluntarily report AI incidents, but there is limited coordinated reporting infrastructure.

PRACTICE 4

Share user feedback

Sharing feedback received on the model through a provider's feedback mechanism

PROCESSES		LEVEL OF ADOPTION (NOV. 2024)	OPEN MODEL CHALLENGES	REQUIRES MODEL-LEVEL COLLABORATION	OTHER RELEVANT STAKEHOLDERS
4.1	Disclose the process of having a feedback mechanism for stakeholders	HIGH		●	**R**
4.2	Aggregate individual user feedback records to provide summaries	MEDIUM	●	●	
4.3	Disclose the feedback follow-up process or, if warranted, the redress mechanism process	LOW		●	
4.4	Create incentive structures to invite stakeholders to participate in the feedback process proactively	LOW		●	

KEY

CCD	Compute, Cloud or Data Providers
CS	Civil Society and Watchdogs
A	Academia
R	Regulators and Norm Enforcers

KEY FINDINGS

- There are feedback mechanisms for users across model providers, such as providing email addresses, feedback forms, and community portals.
- Feedback collections can be seen on community or developer portals, but proactive summaries from model providers are not typically seen.
- There are few incentive structures for providing feedback, though bug bounties act as incentives for highlighting specific issues. It is possible that there are other incentive structures, however, they are not disclosed publicly.
- There is low disclosure of response or redress mechanisms after providing feedback. At the moment, it is possible that there are no disclosure mechanisms that exist, or the mechanisms that exist are not disclosed publicly.

Overall findings

Substantial work is needed from policymakers and actors in the foundation model value chain to ensure that practices related to the documentation of post-deployment impacts are developed and adopted to effectively promote safety and accountability.

The monitoring and documentation of post-deployment impacts for foundation models is an immature field, and significant progress was only evidenced where actors in the value chain had business incentives to adopt practices, such as requesting feedback for model training purposes. Policymakers have not yet put in place the structures to facilitate coordinated, responsible information sharing, and may need to develop rules to incentivize adoption.

While there are processes in place to collect — and potentially assess — information on post-deployment impacts, there is little public or restricted disclosure of this information.

Model providers have processes in place to monitor usage for signs of policy violations and incidents, but there is limited disclosure about findings from these processes. There are examples that model providers can build on, such as Anthropic's transparency around the Clio tool and usage analysis.

Structural processes to enable and incentivize documentation for post-deployment impacts are not yet in place.

Where there is historical precedence for the need of a central structure for post-deployment practices, such as for incident reporting and usage databases, there have only been ad hoc attempts to do so for foundation models (and AI more broadly).

Documenting the impact of open models is in its early stages.

PAI's Risk Mitigation Strategies for the Open Foundation Model Value Chain highlights 15 risk mitigations for app developers and stakeholders at the model-level, and describes challenges for open model governance. We find that there is little adoption of the post-deployment practices described for open models at this early stage, such as for incident and transparency reporting. We explore the substantial challenges in adoption in Section 3.

There are signs that actors are making progress, but there is little detail on exactly how.

Incident reporting is a focus for emerging policy initiatives, as described in PAI's recent Policy Alignment on AI Transparency report, and some organizations have stated in responsible scaling policies that monitoring will play a larger focus moving forward. However, there are few proposals for how exactly this will be implemented. Recent reporting from Anthropic provides an example to build upon for aggregated usage reporting.

SECTION 3

Challenges

Adopting post-deployment practices in this emerging area of governance requires stakeholder education and coordination, and building on precedents from other sectors. This raises several challenges, which are grouped into the following five themes:

CHALLENGE 1

Lack of standardized documentation norms

1.A. Lack of standardization in definitions and documentation requirements

What qualifies as a "serious incident?" What post-deployment environmental information should be shared? In what format should aggregated feedback be shared, and how? This work identified many open questions related to norms for documenting post-deployment impacts, including around definitions and documentation requirements.

Research findings and precedents from other industries that can contribute to post-deployment standardization are highlighted throughout this report, and there have been initial attempts to standardize some definitions and practices. These include OECD's Defining AI Incidents paper and NIST's Managing Misuse Risk for Dual-Use Foundation Models draft report. However, these definitions and practices have yet to be widely adopted — in contrast to pre-deployment documentation where model cards are well-established — and do not always account for open models. This highlights a greater need for effort from policymakers, norm-setters, and standards bodies to "legitimize" these practices through established mechanisms, such as policies and standards.

1.B. Lack of agreed responsibilities

In addition to the lack of standardized definitions and requirements, a common issue described throughout this work was "finger-pointing" — without a common agreement on who should be doing what and limited accountability structures to promote the adoption of practices, it's easy to shift both blame and responsibility toward other organizations. This is exacerbated by the number of stakeholders involved in documenting post-deployment information, especially for open models, and the different model release strategies at play. This vacuum in the definition of responsibilities means that limited progress has been made where post-deployment practices are not aligned with organizational incentives, and in some cases, stakeholders may not be aware of practices related to post-deployment impacts. A greater focus on agreeing on the "taxonomy of responsibilities" for different stakeholders across various release strategies to enable the documentation of post-deployment impacts can help overcome this challenge.

1.C. International interoperability

The use of foundation models by stakeholders worldwide complicates this challenge due to varying and sometimes conflicting regional regulations for transparency and documentation. While international bodies like the UN and OECD aid in setting international AI governance policies, the reality is that different countries have varying levels of AI policy maturity, regulatory definitions, and cultural expectations around transparency and documentation. This creates additional challenges for model providers, model adaptors, and application developers who operate in multiple regions to maintain consistent yet locally interoperable documentation practices. PAI's Policy Alignment on AI Transparency report considers how stakeholders can improve interoperability between leading frameworks.

CHALLENGE 2
Data sharing and coordination barriers

Even if different organizations knew what information to share and their responsibilities in analyzing and documenting it, the current landscape presents operational challenges that act as barriers to implementation. Model providers create foundation models but may lack access to deployment data; app developers generate data but may be unable to share it due to privacy restrictions and competitive concerns; users generate impact data but have privacy protections; and researchers and policymakers need comprehensive data but may be unable to access it.

When data needs to be shared between stakeholders, data privacy is a core challenge — activity logs may contain personal data which cannot legally be shared without consent, and privacy commitments to enterprise customers may block sharing information on use cases or customizations. Privacy-preserving mechanisms exist — such as technical techniques like differential privacy, federated statistics, and differential private federated statistics — and the benefits and limitations of these are explored in PAI's Eyes Off My Data. Safe harbor provisions and vetted researcher access procedures may be adapted for good faith analysis. However, it remains unclear how these techniques and procedures should be applied to post-deployment information analysis, which may require standards development, and organizations are not incentivized to pursue this sort of data sharing (see Challenge 3). Policymakers may need to define requirements to overcome this challenge. There are also contractual barriers when organizations need to share data. A model host or distribution platform may be well-placed to coordinate feedback on models, but the B2B agreements may not be in place to facilitate this legally. Additionally, even in a single organization, privacy and antitrust concerns may restrict data sharing.

Compared to technical documentation about the model's design, development, and deployment, the increased number and variety of stakeholders involved in documentation for post-deployment impacts present a significant coordination barrier. Disclosing an incident — such as a bias and discrimination or security-related harm — may require (1) the user to identify and document the issue and communicate this to (2) the app developer, who then verifies and shares the report with (3) the model provider, who then may need to analyze the report and the model in question and communicate this further internally. This challenge tends to be more acute for open models since their distribution is typically widespread and their monitoring is more difficult. Regardless of how a model is released, stakeholders involved in the deployment of a model should clarify the communication reporting chain and assess compatibility among their processes. Without this, systemic harms — such as bias, discrimination, privacy, and security harms — that impact real users may continue to propagate.

This challenge is even more complex due to the need for interoperable global data privacy and protection regulations. While international organizations like the OECD work to establish common frameworks, regional regulations like the EU's GDPR impose strict requirements on international data transfers. In areas where these regulations are nonexistent or minimal, model-level stakeholders must be mindful of responsibly collecting and utilizing post-deployment impact data so they do not take advantage, misrepresent, or obscure data on marginalized global communities.

CHALLENGE 3
Misaligned incentives

Even when organizations know what to document and how, market incentives may block organizations from sharing this information or encourage other behaviors that are against public interest.

There are incentives for "race to the bottom" behaviors, where capturing market share and user attention is seen as more important than safety research and guardrail development. Rapid deployment may be favored over responsible governance and documentation due to these competitive pressures,[12] which may also motivate organizations to keep market sensitive details private — for example, model providers may not want usage data to be available to competitors. Transparency may also expose organizations to legal scrutiny.

Private stakeholders and public policymakers are incentivized to encourage innovation by protecting commercially sensitive details or research and may also view reducing documentation burdens on companies as a way to promote innovation. However, greater transparency can contribute to innovation by enabling analysis that can reduce societal harms, and by driving accountability for actors in the foundation model value chain. Trust developed through a regulated system can improve investment, adoption, and drive societal benefits.

Sharing aggregated data or information with a trusted third party can also help overcome these innovation incentives, though a lack of precedent and varying regional norms in the field may block wider adoption. This is especially problematic in emerging AI innovation hubs that lack the governance frameworks, technical expertise, and resources needed to implement robust documentation practices for post-deployment impacts, creating significant gaps in understanding global model impacts.

CHALLENGE 4
Limited infrastructure

For other governance activities, a lack of coordination and structure may be an issue, and there is currently no established infrastructure to support the collation and analysis of post-deployment information. For example, the National Highway Traffic Safety Administration's Crash Report Sampling System collects samples on police-reported crashes to form the basis for cost and benefit analyses of highway safety initiatives and regulations, and Cifas manages the "largest database of instances of fraudulent conduct in the UK."[13] However, there is no agreed-upon organization, process, or database that can be used to analyze post-deployment information for foundation models, hindering the adoption of documentation practices for post-deployment impacts.

CHALLENGE 5
Decentralized nature of open model deployment

While open models offer significant benefits, such as increased accountability, innovation, competition, and enabling critical safety research, the decentralized nature of open models presents unique challenges and exacerbates existing barriers to collecting post-deployment impact information, as highlighted in PAI's Risk Mitigation Strategies for the Open Foundation Model Value Chain. While restricted access models allow direct monitoring through APIs, open models require different approaches since they can be run locally and fine-tuned by downstream developers. It is important to note that restricted access models are not immune to misuse or reckless use — they have similar challenges. Restricted access model providers just have more direct levers to monitor and moderate usage.

There are no agreed-upon responsibilities for these approaches, and some data collection mechanisms, such as usage log monitoring, are not currently technically feasible for open model providers. Further study is required on how to balance data collecting and sharing responsibilities, individual privacy and proprietary protections, and security concerns. These challenges affect various actors in the value chain, from model providers to application developers, and impacts the trust built between them.

How these challenges impact practices

These challenges hinder the adoption of these practices, though should not be seen as impossible to overcome. By mapping the challenges to the processes identified, we can identify a route to advancing the field.

KEY TO CHALLENGES

1 Lack of standardized norms
2 Data sharing and coordination barriers
3 Misaligned incentives
4 Limited infrastructure
5 Open models

TABLE 3. **Analysis of which challenges were identified as important for each process and the level of adoption of each process as of Nov. 2024**

			CHALLENGES PRESENT					LEVEL OF ADOPTION (NOV. 2024)
	PROCESSES		1	2	3	4	5	
PRACTICE 1	1.1	Conduct surveys or user research to understand downstream usage						N/A
	1.2	Create tools to support the sharing of activity logs with trusted third parties for analysis	●	●				NONE
	1.3	Implement and track watermarking or identifiers	●				●	NONE
	1.4	Report aggregate usage statistics, across geographies, sectors, or use cases, including usage in high-risk use cases	●	●	●		●	LOW
	1.5	Share information on downstream applications of the model	●	●	●		●	LOW
PRACTICE 2	2.1	Report on labor impact indicators	●	●	●		●	LOW
	2.2	Report on environmental impact indicators	●		●		●	LOW [A]
	2.3	Report on synthetic content impact indicators	●				●	LOW [B]
	2.4	Disclose third-party research access					●	MEDIUM
	2.5	Disclose organizational resourcing commitments and dedicate funding commitments towards post-deployment societal impacts						HIGH
PRACTICE 3	3.1	Monitor for incidents					●	MEDIUM [C]
	3.2	Monitor for policy violations					●	MEDIUM [C]
	3.3	Share summaries of internal incident and policy violation reports	●		●		●	LOW
	3.4	Systematically report AI incidents to a third party	●			●		NONE
PRACTICE 4	4.1	Disclose the process of having a feedback mechanism for stakeholders						HIGH
	4.2	Aggregate individual user feedback records to have as summaries					●	MEDIUM
	4.3	Disclose the feedback follow-up process or, if warranted, the redress mechanism process						LOW
	4.4	Create incentive structures to invite stakeholders to participate in the feedback process proactively						LOW

A Compute information for a majority of models is generally available on established transparency artifacts like model cards and system cards.

B Low indicator here may dictate that not all models have the capacity to produce synthetic content.

C Adoption was assessed across these two processes, as "abuse monitoring" could cover both.

SECTION 4

Recommendations

Considering the field's progress on documenting post-deployment impacts and recognizing the challenges described, we provide the following recommendations:

GENERAL RECOMMENDATION

Implement processes that support sharing usage and feedback information, conducting societal impact research, and reporting incidents and policy violations

KEY ACTIONS

- Model-level and application-level stakeholders should publish available information on violations, incidents, usage, activity logs, feedback, and environmental impacts, in an aggregated format, publicly, and with trusted third parties.
- Organizations and governments should fund and conduct societal impact research, including independent research.

Addressing Challenges 1–5 will contribute to improving the documentation of post-deployment impacts. However, model- and application-level organizations can be more transparent through documentation now.

For example, model providers that also act as application developers should share aggregated information on violation data, usage, feedback, and environmental impact where possible. This is in line with guidelines from PAI's Guidance for Safe Foundation Model Deployment, Risk Mitigation Strategies for the Open Foundation Model Value Chain work, and policy frameworks such as the EU AI Act[A] and supporting General-Purpose AI Code of Practice.[B] Current monitoring practices for restricted access models could be expanded to aggregate and document policy violations and incidents. Anthropic describes how they analyze usage data in a privacy-preserving way and shares insights on usage and activity logs, which could be built on by other model providers. Feedback could be aggregated, analyzed, and made easily accessible. The implementation and adoption of these practices should be tailored to the release strategies and capability of models, with higher capability models adopting more comprehensive documentation in line with their increased risk.[C]

There may be differences for open models. Application-level stakeholders and hosting platforms may be well-positioned to share post-deployment information, as open-model providers may lack direct access to this data. However, high-capability open model providers can explore how digital signatures or "fingerprints" can support monitoring of these models after release. While standardized templates for disclosing this information do not yet exist, early documentation efforts will help shape their development.

[A] Namely EU AI Act Article 13, Article 55, Article 73, and Article 95.

[B] Namely Second Draft General-Purpose AI Code of Practice Commitment 1 and Commitment 17.

[C] While "capability" is an indicator of risk, it is not a direct correlation, and some risks, such as bias, may be exacerbated in lower capability models.

Beyond operational post-deployment impact information, we encourage broader research into these models' economic, environmental, and societal impacts, particularly from academia and civil society organizations with model provider support. Model providers should continue to commit to funding and resourcing comprehensive sociotechnical research initiatives. These commitments should come from internal research efforts, aligned with voluntary commitments coordinated by national governments, such as those coordinated by the US government, and international commitments and directives, such as the G7's Hiroshima AI Principles and Code of Conduct. Support should extend externally to independent research from civil society organizations and other academic institutions, which may look like dedicated long-term funding opportunities and clear protocols for external model and data access. Governments are also well-positioned to fund or conduct broader impact research, learning from examples from Ofcom, the U.S. Bureau of Labor Statistics, and the U.S. Environmental Protection Agency.

RECOMMENDATION 1

Define norms for the documentation of post-deployment impacts through multistakeholder processes, which may be formalized through technical standards

Addresses Challenge 1: Lack of norms and responsibilities

KEY ACTIONS

- Stakeholders should conduct research into methods for documenting post-deployment impacts.
- Stakeholders should contribute to multistakeholder standards development processes.
- Stakeholders should encourage interoperability between standards and policy.

Without a shared understanding of how to implement these practices and who should be responsible for what, there will continue to be "finger pointing" and delays in adopting these practices. The development of post-deployment impact norms, such as best practices, processes, and templates underpinned by shared definitions, should be multistakeholder and inclusive, in line with PAI's Guidelines for Participatory and Inclusive AI.

Some areas require foundational research before norms and standards can be developed, building on the International AI Safety Report and this paper's referenced literature. Two areas that may require additional research include processes **1.3: Implement and track watermarking or identifiers** and **2.3: Report on synthetic content impact indicators**.

Other processes have bases of research that can be used to inform standardization, or have begun to be standardized through unofficial bodies. This might include processes **1.4: Report aggregate usage statistics, across geographies, sectors, or use cases, including usage in high-risk use cases**, where there is academic literature on information that could be included, and **3.4: Systematically report AI incidents to a third party**, where OECD's Defining AI Incidents paper provides a basis for standardizing incident terminology.[14]

Forums for standardization include international standard-setting bodies, such as ISO/IEC and IEEE, and national or regional standardization/quasi-standardization initiatives, such as NIST's AI RMF and CEN/CENELEC.

Interoperability should be a key consideration in developing these standards, as discussed in PAI's Policy Alignment on AI Transparency report. Standards for documenting post-deployment impacts should build on existing definitions and policies highlighted in this report, such as NIST's Managing Misuse Risk for Dual-Use Foundation Models draft report.

RECOMMENDATION 2

Explore mechanisms for responsible data sharing.

Addresses Challenge 2: Data sharing and coordination barriers

KEY ACTIONS

- Model- and application-level stakeholders should explore how technical privacy-preserving mechanisms can be applied to sharing documentation on post-deployment impacts.
- Model- and application-level stakeholders should foster environments for sharing information for disclosure.
- Model- and application-level stakeholders should explore options for academic and civil society data access procedures.

Once responsibilities and processes are in place to collect information on post-deployment impacts, they may need to be shared between stakeholders to enable analysis while preserving user privacy and business sensitive details. Responsible data sharing will require technical privacy-preserving mechanisms, an environment that facilitates data sharing in the ecosystem, and specific procedures for academic and researcher access. However, it is important to note that there cannot be responsible data sharing without responsible data collection, as elaborated in PAI's Guidelines for Participatory and Inclusive AI.

Privacy-preserving mechanisms, such as differential privacy, federated statistics, and differential federated privacy, can enable valuable research and analysis while preserving user privacy. Model providers, model hubs, application developers, distribution platforms, academia, and civil society may be well-placed to explore how to apply these mechanisms to post-deployment data sharing. This could build on previous research applicable to AI, such as PAI's Eyes Off My Data, and other domains, such as healthcare and census data collection. Anthropic's Clio and Economic Index also provides an example of how restricted access model providers can analyze and share aggregated usage insights on use cases in a privacy-preserving manner.

As post-deployment impact information is distributed throughout the foundation model value chain, model- and application-level stakeholders should share information, such as usage domain and use case, with model providers for the purpose of government disclosure or public release. This may involve incorporating data sharing in contractual discussions, with the stated purpose of disclosing for accountability. This may also be coordinated through a third party, drawing lessons from FDA post-deployment monitoring.

As highlighted in Practice 2, academia and civil society stakeholders play essential roles in understanding the societal impacts of foundation models. However, these researchers often lack access to essential information held by model- and application-level stakeholders. Therefore, we suggest that data sharing mechanisms should account for the stakeholder power dynamics at play and ensure equitable access for academic institutions and researchers and civil society researchers to conduct independent research for the sake of public interest. Research calls for legal and technical safe harbor provisions to "[indemnify] public interest safety research and [protect] it from the threat of account suspensions or legal reprisal." Safe harbors are also highlighted in NIST's Managing Misuse Risk for Dual-Use Foundation Models draft report. These rules should be developed in a multistakeholder forum and build on previous research aimed at evaluation and lessons learned from EU data altruism initiatives.

The EU is developing rules for data access for researchers through the Digital Services Act for platforms and search engines, which aims to "provide access to data for the purpose of conducting research that contributes to the detection, identification, and understanding of systemic risks." This approach could be explored to better understand the impact of foundation models post-deployment. Another option is to develop tools that allow users to give consent for their data to be used in valuable research datasets, such as WildChat.

RECOMMENDATION 3
Policymakers should explore where guidance and rules on documenting post-deployment impacts are needed

Addresses Challenge 3: Misaligned incentives

KEY ACTIONS

- Policymakers should review how documentation on post-deployment impacts can help them identify and assess risks.
- Policymakers should identify where the current levels of adoption in documenting post-deployment impacts conflict with policy objectives.
- Policymakers should explore how guidance and rules for monitoring and documenting post-deployment impacts can align with the EU Code of Practice.[C]

Policymakers, especially national and regional legislators, play an important role in counterbalancing market incentives that might lead industry to act against the public interest. The EU's transparency reporting requirements for online platforms and search engines demonstrate how policymaker intervention can successfully mandate documentation of post-deployment impacts, such as country-specific usage data, and similar interventions may be beneficial in this domain.

Firstly, policymakers should review how information on post-deployment impacts can help them to assess risks that may impact their economic, societal, or environmental policy

[C] Only the second draft of the General-Purpose AI Code of Practice is currently available, with future iterations planned for release.

objectives. This information is important for effectively assessing what benefits and harms are being caused, and policymakers may be able to conduct or fund research to better understand these impacts, as discussed in the General Recommendation.

Policymakers should also identify where current levels of adoption in the practices laid out in this report are hindering policy objectives. For example, a lack of voluntary incident reporting may conflict with safety-related policy objectives, so voluntary or binding rules and frameworks could encourage adoption. This process should consider the urgency and importance of these behaviors, the technical complexity required for rules development, the level of independence required to set rules, and other relevant factors.

One of the most influential and detailed policy initiatives is the EU AI Act, which lays out binding requirements for foundation model providers, and describes in detail how providers can meet those requirements through the Code of Practice.[D] As noted in PAI's Policy Alignment on AI Transparency report, interoperability across jurisdictions may be beneficial, so policymakers should explore how to align their frameworks and rules with the EU Code of Practice and other emerging standards and legislation. For example, the second draft of the General-Purpose AI Code of Practice details monitoring practices that should support risk assessment, documentation that should be disclosed to downstream developers, and documentation disclosed to the AI Office through a Safety and Security Framework — policymakers could look to these practices for policy development guidance.

RECOMMENDATION 4

Policymakers should develop blueprints for national post-deployment monitoring functions

Addresses Challenge 4: Limited infrastructure

KEY ACTIONS

- Policymakers should commit to investing in the capacity, capabilities, and structures that will enable post-deployment monitoring of foundation models by September 2025.
- Policymakers should develop blueprints for post-deployment incident monitoring of foundation models by March 2026.
- Policymakers should build on incident reporting structures to monitor usage of foundation models that exceed agreed-upon risk or capability thresholds by March 2027.

Though there are third-party incident databases, such as the OECD AI Incidents Monitor and AI Incident Database, the foundation model industry lacks a database with industry and governmental buy-in for tracking incidents and usage, such as the CVE Program for cybersecurity or the National Highway Traffic Safety Administration's Crash Report Sampling System. Policymakers should define what a system that achieves the same goals for foundation models should look like — i.e., a blueprint. This should involve a review of potential options for foundation model incident monitoring functions, and could build on the International Network of AI Safety Institutes, which provides the prerequisite structures

[D] The EU AI Act uses the term "General Purpose AI Providers" instead of "Foundation Model Providers."

and relationships with model providers, or the pre-established OECD AI Incidents Monitor. This may be for foundation models over a certain threshold, as discussed in PAI's Policy Alignment on AI Transparency report. This may build on analysis from other domains, proposals for AI Model Registries[E] and existing, but undetailed, policy initiatives[15] and legislation.[16] It should look to feed into international discussions, such as the proposed UN policy dialogues. Whistleblower protections and legal safeguards may be required to ensure that stakeholders can report incidents without fear of reprisal.

While developing a full blueprint may take time, the enabling factors for this can be committed soon. Many jurisdictions have already highlighted incident reporting as an important governance measure, so resource commitments are a logical next step.[F]

The same capabilities and structures built through this blueprint, such as relationships with model providers, secure database management and analysis functions, will also support the collection and analysis of usage information. This may not be prioritized by model providers and deployers, but provides significant societal benefit so policymakers are well placed to intervene. High-risk sectors should be prioritized instead.

Policymakers in jurisdictions with pre-existing AI functions, such as AI Safety Institutes, may be best placed to develop blueprints at this stage. However, given the importance of incident reporting to identifying post-deployment impacts, all jurisdictions should explore ways to achieve incident reporting-related policy goals.

RECOMMENDATION 5

Conduct research into methods for collecting information on open model impacts

Addresses Challenge 5: Decentralized nature of open models

KEY ACTIONS

- Open model providers should collaborate with academia, civil society, and other open model providers to conduct research into the methods for collecting usage information and identifying and sharing societal impact indicators.
- Open model providers should collaborate with application developers and model-hosting services to determine realistic responsibilities for monitoring incidents and policy violations for open models.

Since the documentation of post-deployment impacts of open models requires more stakeholder coordination and presents unique challenges compared to restricted access models, open model providers should research how to responsibly collect information on post-deployment impacts while addressing technical feasibility and monitoring challenges.

Research could be targeted around the options laid out in the PAI's Risk Mitigation Strategies for the Open Foundation Model Value Chain, which include:

[E] For example, CSET's An Argument for Hybrid AI Incident Reporting.

[F] See analysis in PAI's Policy Alignment on AI Transparency report.

- **Release models with digital signatures or “fingerprints’.**[17] This helps track the model’s provenance and its outputs to allow insights into a model’s traceability and usage information (related to Practice 1 and Practice 2).
- **Implement disclosure mechnisms, such as watermarking, for AI-generated content.**[18] This helps determine impact indicators around synthetic media impacts, which can be either beneficial or harmful (related to Practice 2).
- **Develop and implement durable model-level safeguards.**[19] Specifically, more research should be invested in pre-training models with difficult-to-remove safety mechanisms, such as self-destructing models that break when users attempt to alter or remove safety guardrails (related to Practice 3 around monitoring for policy violations).
- **Monitor misuses, unintended uses, and user feedback.**[20] This requires shared responsibility between an open model provider and an application developer since developers may have more control and visibility over how their applications are being used, making monitoring for misuses and unintended consequences easier (related to Practice 3 and Practice 4).

Open model providers should adopt a multistakeholder approach and do not have to do this research alone, as academia, civil society organizations, and other open model providers may want to collaborate and address these challenges too. Academia and civil society organizations may also have more insights into the tradeoffs between user privacy concerns, security, and trust-building that are further exacerbated by open models. Shared responsibilities and accountability are necessary for open models since they involve so many stakeholders.

Key questions

Beyond these recommendations, there are key questions that we encourage further research and discussion into:

- For which issues, whether due to urgency or public interest, should policymakers lead the definition of best practices and develop binding rules?
- How can specific processes be implemented for open models?
- What additional practices for documenting post-deployment impacts are important to foundation model governance?
- What level of detail is required in documentation to measure relevant impacts and assess trustworthiness?
- What counts as a substantial modification to a model? When an organization substantially modifies a model, what are its responsibilities for documenting post-deployment impacts, and how can this administrative burden be managed?

PAI will continue to improve collective understanding of the field and drive accountability through future progress reports. If you would like to know more, please contact policy@partnershiponai.org.

APPENDIX 1

Methodology

This report was developed taking the following steps:

1. PAI established the need for a "progress report" to assess progress and accountability related to specific practices with global experts in our Policy Steering Committee. We selected the documentation of post-deployment impacts as a key area to explore that bridges potential gaps in public policy and practice.
2. A focused working group on the documentation of post-deployment impacts was established with 18 organizations, including partners engaged in PAI's ABOUT ML program and Policy Steering Committee.
3. PAI defined the goals and driving questions related to the documentation of post-deployment impacts and identified specific practices to explore, building on PAI's Guidance for Safe Foundation Model Deployment.
4. PAI shared an initial questionnaire to inform discussions about the benefits and challenges of post-deployment practices and received 13 responses.
5. PAI held four working group sessions to refine the approach to discussing practices, benefits, and challenges.
6. PAI held a workshop with industry, civil society, and academia to discuss the benefits, challenges, and potential responsibilities of different stakeholders for four practices related to documenting post-deployment impacts.
7. PAI conducted an analysis of workshop findings, supported by an additional literature review, to provide a draft of the report.
8. Contributors reviewed the draft report, with information shared asynchronously and through two additional working group sessions, and provided over 100 comments and suggestions.
9. PAI reviewed each comment and edited the report accordingly. While all comments were reviewed and the vast majority were accepted, PAI retained authorship of the report and did not take on all comments and suggestions.

Understanding the current state

There are various deployment configurations for foundation models, which leads to complications in measuring and defining responsibilities for each post-deployment practice.

Though the configuration of stakeholders will impact the ease and method of implementation of the practices, post-deployment documentation is vital to ensuring the benefits laid out in this report. Therefore, to measure the progress of the field, we primarily focus on the actions of model providers and take the following approach:[A]

[A] The examples and analysis from steps 1–3 is shared in this Google Sheets document.

1. **Identify 12 model providers that provide a "representative" reflection of the field** (see Limitation #4 for information)

OPEN MODELS	RESTRICTED ACCESS MODELS
Allen Institute for AI: OLMo	**Anthropic:** Claude 3.5
BigCode: StarCoder	**Cohere:** Command
IBM: Granite	**Google:** Gemini 1.5
Meta: Llama 3.1	**Inflection:** Inflection 2.5
Mistral: Mistral Small	**OpenAI:** GPT-4
Microsoft: Phi-3	
Stability AI: Stable Diffusion 3	

These models are primarily created in the US and EU, though their reach and impact affect a global audience.

We describe the "level of adoption" of processes where model providers are key actors, and also highlight supporting actors. To assess the "level of adoption" of a process, we use the following guide:

2. **Conduct a review of publicly available information to assess whether each provider has adopted, or is collaborating to adopt the stated process.** Where the process is not related to the actions of a model provider, we have reviewed publicly available information and provided a written description of the progress of the field.
3. **Ask model providers to share additional information.** We reached out to each model provider to confirm that the review was factually accurate and to share additional information.[B]
4. **Display the number of providers that have adopted this process and rate the level of adoption of the processes,** using the following rating scale:

LEVEL OF ADOPTION[C] (NOV. 2024)	PRACTICES IDENTIFIED (OUT OF 12)	INTERPRETATION
NONE	0	No organizations have implemented this process. Significant work is required to overcome the challenges blocking adoption.
LOW	1 — 3 *(or >0 partial implementations)*	Few organizations (≤25%) have implemented this process, or some organizations have partially implemented it. The field can build on these initial practices, but more work is needed.
MEDIUM	4 — 6	Some organizations (26%-50%) have implemented this process. The field can learn from the good practices highlighted.
HIGH	7 — 12	Most organizations (>50%) have implemented this process. Future work should focus on identifying and aligning around best practices.

For one process, model providers may not play a key role in the process, so we do not measure the level of adoption and highlight this using ***N/A****.*

[B] We were unable to get responses from Inflection, Mistral, OpenAI, and Stability AI.

[C] We do not assess the quality of the adopted practices when describing the "level of adoption," but do explore them in more detail in the Annex.

Limitations

We recognize the limitations of this approach, and have taken the following mitigation steps to minimize their impact.

LIMITATION	DESCRIPTION	MITIGATION
1. Not all of the processes highlighted relate to model providers or model-level organizations.	This work is based on the Model Deployment Guidance, which focuses on the actions of model deployers. However, some of the processes highlighted describe the actions of stakeholders who are not at the model level, but are still important for the effective deployment of this practice.	We have highlighted where this is the case with "N/A."
2. We do not assess the quality of the processes.	Some of the processes may have the same description of their implementation, but actually differ from each other significantly. For example, two different model deployers may both say they conduct monitoring for policy violations, but may employ different algorithms to flag these violations, which may have different success or error rates.	We provide written discussions of the progress of each process to highlight these nuances.
3. There are variations between open and closed model practices.	It's acknowledged that open models and restricted access models require different governance and documentation practices.	We provide written discussions of the nuances for open and closed models for each process to highlight these nuances. We highlight processes with significant challenges for open models.
4. We are not assessing the entire field of stakeholders.	Though we aim to discuss the progress of "the field," we can not feasibly assess all the stakeholders who have developed or interacted with foundation models.	We select a representative sample of models, choosing models with a variety of "openness." We highlight global implications throughout the report.
5. "Yes or no" scoring systems may not reflect the nuances of the subject matter.	While assessing each process, we make a decision on whether a specific implementation of the process is sufficient. This may mean that a partial implementation, such as providing total model usage but not use case and geography usage, is not fully accounted for in the rating.	We use a "none, low, medium, high" rating system to ensure that the focus is on the overall progress of the field. We provide substantial written context to each process.

APPENDIX 2

Reference information

Risks that post-deployment documentation can help to manage

We identified the following risks as common to multiple risk frameworks, and refer to them throughout this report to emphasize how post-deployment information can support risk managment.

TABLE 4. **Risks common to the EU AI Act Code of Practice process, NIST AI RMF, AI Risk Repository and PAI's Guidance for Safe Foundation Model Deployment**

RISK CATEGORY	GENERATED CONTENT MIGHT...	EXAMPLES
Discrimination, Toxicity & Bias	Unfairly represent certain groups or individuals	Harmful decision-making, toxic or hate speech
Privacy & Security	Leak and have unauthorized disclosure or de-anonymization of sensitive data. May also expose model vulnerabilities	Compromised biometric, health, location, or personally identifiable data
Information Integrity	Have false or misleading information	Misinformation, dangerous or violent information
Malicious Uses	Be used by bad actors	Chemical, biological, radiological, and nuclear risks (CBRN), Non-consensual intimate imagery (NCII)
Human-Computer Interaction (HCI)	Impact the user's emotional and physical well-being	Anthropomorphization, or emotional entanglement between humans and AI systems
Societal Harms	Have large negative consequences on society as a whole	Increase inequality, environmental harms, and labor harms
AI Systems Safety, Failures & Limitations	Be due to technical faults and limitations of a model or system	Hallucinations/fabrications, emergent capabilities, loss of control

Considerations in documenting post-deployment impacts

- **The model release type affects how information can be collected and which stakeholders should be involved.** See Table 1 and Challenge 5 for more information.
- **Not all information should be disclosed immediately and publicly.** There is a significant benefit to sharing information openly, but some information should only be disclosed to trusted stakeholders for privacy, security, or other reasons. For example, publicly sharing a full incident report on a user data vulnerability may violate user privacy rights so a carefully redacted disclosure may be more appropriate.
- **The model's capability[A] will affect the information that should be disclosed.** Good governance practices involve imposing practices that are proportionate to the risk. Models with higher capabilities may pose higher or "systemic" risks that may require more stakeholders to help mitigate. This approach is born out in PAI's Guidance for Safe Foundation Model Deployment and multiple policy frameworks.[21] However, risks do not entirely scale with capability due to some risks, such as bias and discrimination, being more severe in less capable models.
- **Information shared privately or publicly should focus on audience needs and present this information accordingly.** To be effective, stakeholders developing documentation should understand audience needs and ensure that it is legible for that audience.
- **Disclosure of this information may inform changes in practice.** The purpose of documentation is not to simply document information. Insights gained by model providers when documenting and reflecting on information, such as when reviewing feedback or impact information, may inform changes in practice. More formally, regulation may require mitigations where harms or issues are identified. In either case, documentation should be actionable in that it "contains the appropriate level of granularity and detail to enable informed decision-making for its intended audience."[22]
- **Foundation models will be fine-tuned by downstream stakeholders, and it can be difficult to define where responsibility for negative impacts lies.** While the duty of model providers' is to prevent foreseeable harms, the adaptation of a model will change how it functions. It's also possible that policy initiatives focused on pre-deployment documentation to ensure that actions taken before any adaptation are responsible. Responsibilities for some information elements will lie with different actors in the value chain, and collaboration between actors will be necessary to understand some of the downstream impacts.
- **Some impact information is difficult to gather on a short- to medium-term timeline.** Broader economic and societal impacts may take years to become measurable through specific indicators hence why they are difficult to identify and document.

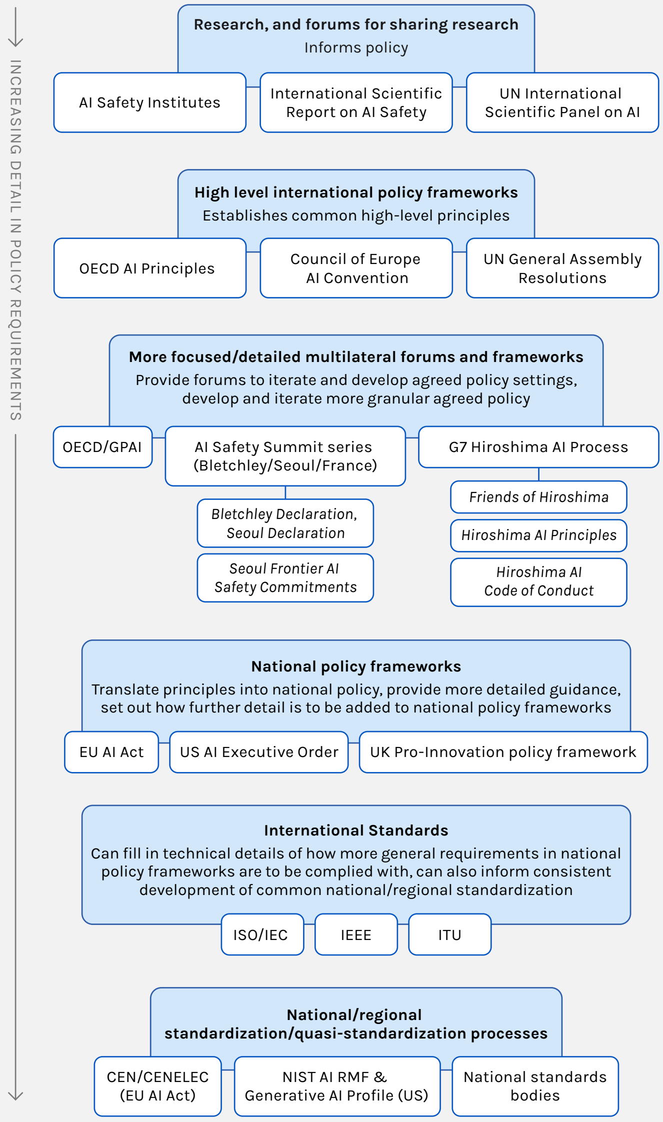
[A] How "capability" should be assessed, and what thresholds should apply to distinguish requirements by "capability," is still a matter of debate in the field.

Initiatives that inform AI policy

This report uses “policy” to refer to voluntary or binding rules and frameworks that are developed by intergovernmental organizations, such as the UN and G7, and regional and national governmental organizations, such as the EU Commission and NIST, to achieve specific outcomes. “Policymakers” refers to the people and organizations that develop policies.

Non-policymakers also inform and contribute to policy development. The diagram at right, explored in PAI’s Policy Alignment on AI Transparency report, provides an overview of how different organizations and initiatives inform the development of policies and legal instruments.

FIG 1. **Informing coherent national policies through the AI governance stack**


INCREASING DETAIL IN POLICY REQUIREMENTS
Research, and forums for sharing research
Informs policy
AI Safety Institutes
International Scientific Report on AI Safety
UN International Scientific Panel on AI
High level international policy frameworks
Establishes common high-level principles
OECD AI Principles
Council of Europe AI Convention
UN General Assembly Resolutions
More focused/detailed multilateral forums and frameworks
Provide forums to iterate and develop agreed policy settings, develop and iterate more granular agreed policy
OECD/GPAI
AI Safety Summit series (Bletchley/Seoul/France)
G7 Hiroshima AI Process
Bletchley Declaration, Seoul Declaration
Seoul Frontier AI Safety Commitments
Friends of Hiroshima
Hiroshima AI Principles
Hiroshima AI Code of Conduct
National policy frameworks
Translate principles into national policy, provide more detailed guidance, set out how further detail is to be added to national policy frameworks
EU AI Act
US AI Executive Order
UK Pro-Innovation policy framework
International Standards
Can fill in technical details of how more general requirements in national policy frameworks are to be complied with, can also inform consistent development of common national/regional standardization
ISO/IEC
IEEE
ITU
National/regional standardization/quasi-standardization processes
CEN/CENELEC (EU AI Act)
NIST AI RMF & Generative AI Profile (US)
National standards bodies

APPENDIX 3

Acknowledgments

This report was prepared with guidance from PAI's Policy Steering Committee.

We appreciate the invaluable input provided by experts who participated as members of PAI's post-deployment documentation working group, engaged in workshop discussions, provided written comments, or otherwise provided input into this report, including:

Working group members who contributed to this work:

Alex Kessler, Microsoft

Amy Smith, Intuit

Alejandro Segarra & Valeria Milanes, Asociación por los Derechos Civiles

Amy Winecoff, Center for Democracy & Technology

Bogdana Rakova, DLA Piper

Connor Dunlop, Ada Lovelace Institute

Connor Wright, Montreal AI Ethics Institute

David Wakeling & Marcus Turner, A&O Shearman

Fion Lee-Madan, Fairly.ai

Gemma Galdon Clavell, Eticas.ai

Jenn Wortman Vaughan, Microsoft

Michael Hind, IBM Research

Maxine Setiawan, Ernst & Young

Michael Kleinman, Meta

Richard Mathenge, African Content Moderators Union

Reena Jana, Google

Shamira Ahmed, Data Economy Policy Hub

Sam Jung, Allen Institute for AI*

Victoria Matthews, Sony AI

Special thanks to the experts who helped shape conversations and produce critical insights from our initial questionnaire and the Policy Forum workshop, in addition to our working group members:

Amanda Craig Deckard, Microsoft

Deon Woods Bell, Gates Foundation

Diane Chang, Cohere

Finale Doshi-Velez, Harvard

Gabriel Nicholas, Center for Democracy & Technology*

Hadrien Pouget, Carnegie Endowment for International Peace*

James Baker, Meta

Kevin Klyman, Stanford HAI

Lama Nachman, Intel

Lisa Pearlman, Apple

Marc Etienne Ouimette, AWS

Raina Wazir, leiwand.ai

Sebastian Hallensleben, CEN/CENELEC and VDE*

Tiffany Georgievski, Sony AI

Will Cutler, UK Government

Special thanks to PAI staff who have been integral to the development of this report:

Aimee Bataclan, John Howell, Madhulika Srikumar, Neil Uhl, Rebecca Finlay, Stephanie Bell, Stephanie Ifayemi, Tina Park

* This person has since moved on from their organization at the time of the report release

ANNEX

Examples of documentation practices

This section highlights examples of implemented practices and processes that were identified throughout this project. This is not a comprehensive list of all practices adopted in the field. A fuller analysis of these examples can be seen in this Google Sheets document.

1. Share usage information

EXAMPLE	SOURCE	DISCUSSION
1.1: Conduct surveys or user research to understand downstream usage		
Online Nation 2023 Report	Ofcom	Provides varied statistics on the usage of ChatGPT.
Stanford HAI Index	Stanford HAI	Highlights the adoption of generative AI in organizations.
Millions of People Are Using Abusive AI 'Nudify' Bots on Telegram	Wired	WIRED reviewed the usage of Telegram bots and found that "more than 4 million 'monthly users'" are using bots that produce explicit nonconsensual content.
1.2: Create tools to support the sharing of activity logs with trusted third parties for analysis		
WildChat	WildChat	WildChat is a research initiative that analyzed 1 million ChatGPT interactions "in the wild".
PySyft	OpenMined	Allows users to conduct data science on non-public information without seeing or obtaining a copy of the data itself.
1.3: Implement and track watermarking or identifiers		
SynthID	Google	SynthID watermarks and identifies AI-generated content by embedding digital watermarks directly into AI-generated images, audio, text or video. This is a tool relevant to watermarking, but does not count as being adopted by a model provider.
1.4: Report aggregate usage statistics, across geography, sector or use case, including usage in high-risk use cases		
Foundation Model Transparency Index disclosures	Stanford CRFM	Google, Microsoft, Mistral and Stability AI explicitly state that they do not share usage data externally. Stability AI share that they might consider "releasing aggregate usage statistics of stablevideo.com on reaching some milestones".
Clio: Privacy-preserving insights into real-world AI use and Economic Index	Anthropic	Clio is an automated analysis tool that enables privacy-preserving analysis of real-world language model use. It gives Anthropic insights into the day-to-day uses of claude.ai in a way that's analogous to tools like Google Trends. Anthropic report initial usage analysis across use cases and language from analyzing 1 million conversations.
Llama usage blog	Meta	Meta provides overall usage of Llama models, in partnership with model hosts: "Hosted Llama usage by token volume across our major cloud service provider partners more than doubled May through July 2024 when we released Llama 3.1."
Hugging Face download statistics (e.g. OLMo 1B July 2024; LLama 3.1 8B)	Hugging Face	Provides total usage by token volume.

1.5: Share information on downstream applications of the model		
Google Cloud's GenAI Use Cases	Google	These examples promote positive use cases.
Gemma use cases		
Search improvements		
Advancing medical AI with Med-Gemini		
Stability AI Customer Stories	Stability AI	
Cohere use cases	Cohere	
Llama usage blog	Meta	
Hugging Face spaces statistics (e.g. Llama 3.1 8B; Stable Diffusion 3 Medium)	Hugging Face	Non-comprehensive view of downstream application statistics.

2. Enable and share research on post-deployment societal impact indicators

EXAMPLE	SOURCE	DISCUSSION
2.1: Reporting on labor impact indicators		
Anthropic's Economic Index	Anthropic	Example of Anthropic, who is a model provider, providing an understanding of AI's effect on the labor market through data and insights from Claude.ai.
Generative AI, the American worker, and the future of work	Brookings, OpenAI	Example of collaboration by Brookings and OpenAI to research job and labor impacts in the United States .
Generally Faster: The Economic Impact of Generative AI	Google, Independent Researcher	Example of independent research being supported by Google around the economic impacts of genAI (a generalization of all models not focused on a model).
Guidelines for AI and Shared Prosperity	PAI	Example of multistakeholder work to provide frameworks and tools to assess job and labor impacts for various stakeholders in the value chain.
Data Enrichment Sourcing Guidelines	PAI	Example of multistakeholder work to provide a framework for model-level stakeholders (Model Providers, Model Adaptors, Model Optimizers) and downstream actors to advance just labor conditions for data enrichment workers.
Implementing Responsible Data Enrichment Practices at an AI Developer	PAI	Example of collaboration by PAI and DeepMind to research the practicality of the Data Enrichment Guidelines for DeepMind's use.
Generative AI's Labor Impacts: A Three-Part Series	Data & Society	Examples of collaborations by Data & Society and various AI Subject stakeholders to discuss the labor impacts on workers.
OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic	Time	Example of watchdog reporting by Time Magazine on OpenAI's labor practices with testimonials from AI Subjects.
Potential Labor Market Impacts of Artificial Intelligence: An Empirical Analysis	White House, Council of Economic Advisors	Example of labor and job impact reporting by an internal Regulator/ Norm Enforcer body. This report extends collaborative engagements with researchers, scholars, and stakeholders to influence policy.
2024 Work Trend Index Annual Report	Microsoft and LinkedIn	Example of Microsoft and LinkedIn reporting on how AI will reshape work and the labor market broadly, surveying 31,000 people across 31 countries, identifying labor and hiring trends from LinkedIn, and analyzing trillions of Microsoft 365 productivity signals as well as research with Fortune 500 customers. This report includes CoPilot as a case study.
The Impact of AI on Developer Productivity: Evidence from GitHub Copilot	Github, Microsoft	Example of a service provider (Github) conducting productivity research their own service.

2.2 Reporting on environmental impact indicators		
Foundation Model Transparency Index disclosures	Stanford CRFM	Only BigCode, IBM, and Meta provided information about compute usage, energy usage, and carbon emissions. BigCode and IBM listed other environmental considerations like water usage and data center impacts in their governance cards. Stability AI disclosed energy usage and carbon emissions. Microsoft only disclosed their compute usage.
The Uneven Distribution of AI's Environmental Impacts	Harvard Business Review	Example of Academia platforming environmental impacts from AI.
AI brings soaring emissions for Google and Microsoft, a major contributor to climate change	NPR	
Microsoft's Hypocrisy on AI	The Atlantic	Example of reporting from a watchdog organization raising alarm about increased emissions due to genAI.
Sustainable AI: Environmental Implications, Challenges and Opportunities	Meta Research	Example of reporting environmental impacts from an internal perspective by Meta research to incentivize other model providers to do the same.
Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model	Hugging Face, Graphcore, LISN & ENSIIE	Example of reporting environmental impacts from an internal perspective by the model provider (Hugging Face) with collaborations from a hardware provider and academia.
2.3: Reporting on synthetic content impact indicators		
Responsible Practices for Synthetic Media: A Framework for Collective Action	PAI	Examples of multistakeholder work to provide a framework for model-level stakeholders (Model Providers, Model Adaptors, Model Optimizers) and application developers on how to develop, create, and share synthetic media responsibly.
Case Studies of real-world examples operationalizing the Responsible Practices for Synthetic Media	PAI	Examples of collaborations between various stakeholders with PAI to highlight areas where synthetic media governance can be applied, augmented, expanded, and refined for use in practice.
How Meta changed its approach to direct disclosure based on user feedback	Meta	
SynthID	Google	SynthID watermarks and identifies AI-generated content by embedding digital watermarks directly into AI-generated images, audio, text or video. This is a tool relevant to watermarking, but does not count as being adopted by a model provider.
Microsoft's Responsible AI Standard	Microsoft	"Microsoft AI systems are designed to inform people that they are interacting with an AI system or using a system that generates or manipulates image, audio, or video content that could falsely appear to be authentic" — Goal T3: Disclosure of AI interaction.
2.4: Disclosures around third-party research access		
Granting Access: Supporting Researchers to Use LLMs	Cohere	Examples of different disclosures around support for third-party research and researchers which includes access, funding, and other forms of support.
Cohere For AI Scholars Program: Research Journeys Start Here	Cohere	
Generally Faster: The Economic Impact of Generative AI	Google, Independent Researcher	
Microsoft Research AI & Society Fellows	Microsoft	
Google PhD fellowship program	Google	
Llama Impact Grants	Meta	

2.5: Disclosures around organizational resourcing commitments and dedicated funding commitments towards post-deployment societal impacts		
The precautionary principle: partnering with the White House on AI safety	Inflection	Examples of model providers signing on to the White House voluntary commitments in 2023. Signatories include Amazon, Anthropic, Google, Inflection, Meta, Microsoft, Apple, Stability AI, Cohere, and OpenAI.
Fact Sheet on voluntary commitments	White House	
Voluntary Commitments by Microsoft to Advance Responsible AI Innovation	Microsoft	
Cohere Joins Enterprise-Focused Cohort on White House AI Commitment	Cohere	
Fulfilling the Voluntary Industry Commitments on AI	Google	
Stability AI Joins IWF's Mission to Make Internet a Safer Space for Children	Stability AI	As an Internet Watch Foundation (IWF) Member, Stability AI now has access to a suite of cutting-edge tools developed to stop the spread of criminal videos and images on the internet, such as the IWF Hash List. The Hash List is a special catalogue of criminal images that have been given individual hashes that are completely unique. A hash is a type of digital fingerprint that identifies a picture of confirmed child sexual abuse. By using the IWF's Hash List, tech companies can stop criminals from uploading, downloading, viewing, sharing or hosting known images and videos showing child sexual abuse on the internet.

3. Report incidents and disclose policy violations

EXAMPLE	SOURCE	DISCUSSION
3.1 and 3.2: Monitoring for incidents and policy violations		
Microsoft Abuse Monitoring	Microsoft	Microsoft detects and mitigate policy violations, but do not disclose summaries of this information. They also classify prompts and completions according to risk categories. Microsoft has established processes to receive vulnerability reports from external finders and work with them through stages of investigation, remediation, and provision of public information (as needed).
Azure OpenAI Service content filtering		
Coordinated Vulnerability Disclosure (CVD) policy		
GPT-4o System Card	OpenAI	OpenAI "enforces [their] Usage Policies through monitoring and take action on violative behavior in both ChatGPT and the API."
Abuse monitoring | Generative AI on Vertex AI	Google Cloud	Google uses automated safety classifiers to detect potential abuse and violations.
Legal terms and conditions	Mistral	Mistral monitors abuse across its services.
Stability AI		Stability AI committed to "detect and remove child safety violative content on your platforms" but no statement on monitoring can be found.
Cohere Trust Center — Monitoring; Cohere Enterprise Data Commitments	Cohere	Cohere conducts various forms of monitoring, and for Enterprise use, Cohere "log[s] and monitor[s] the use of [their] SaaS Platform for compliance with [their] customer agreements, Usage Policy, and for security risks to [their] services."
LlamaGuard	Meta	Meta has developed a tool for downstream use that classifies inputs and responses as unsafe.
Granite Guard	IBM	IBM has developed a set of models that are "designed to detect risks in user prompts and LLM (large language model) responses".
Model Governance — IBM watsonx .governance	IBM	IBM offers a product which enables the monitoring of inputs or outputs for harmful content.

3.3: Sharing summaries of internal incident and policy violation reports		
Responsible Scaling Policy, Version 1.0	Anthropic	In 2023, Anthropic were “preparing” to disclose incidents on specific threats and vulnerabilities with other labs, and committed to maintaining a publicly available channel for privately reporting model vulnerabilities for ASL-3 models. In October 2024, Antropic committed to “periodically release information on internal reports of potential instances of non-compliance”, and publicly disclose summaries of Safeguards reports. Nothing has been disclosed as of November 2024.
Responsible Scaling Policy	Anthropic	
OpenAI Status — Incident History	OpenAI	Details performance and uptime incidents.
Other third party databases		
AI Incident Database	AIID	Users can submit incidents to the AIID. AIID provides regular summaries on incidents reported, and has shared operational insights into the challenges of cataloging AI incidents.[23] Examples include: AI Incident Roundup – August and September 2024 and AI Incident Roundup – July 2024.
OECD AI Incidents Monitor (AIM)	OECD	The stated goal of the AIM is to “track actual AI incidents and hazards in real time and provide the evidence-base to inform the AI incident reporting framework and related AI policy discussions.” Incidents from the media are “identified and classified using machine learning models.”
AI, Algorithmic, and Automation Incidents and Controversies repository	AIAAIC	The AIAAIC is an “independent, open, public interest resource that details incidents and controversies driven by and relating to AI, algorithms and automation.”

4. Share user feedback

EXAMPLE	SOURCE	DISCUSSION
4.1: Disclosing the process of having a feedback mechanism for stakeholders		
Llama Output Feedback	Meta	Allowing users to disclose the prompt given to a model and the response from the model specifically for problematic content related to criminal or regulated activity (like weapons or illegal substances), content you find hateful or harmful (like slurs or bullying).
Send Feedback with Gemini Apps	Google	Allows application developers and other users to provide feedback to model providers through a variety of ways found on this landing page.
IBM watsonx Feedback Portal	IBM	Allows application developers to provide feedback to model providers from a model hub’s portal.
Thumbs Up/Thumbs Down function for feedback within Claude’s UI	Anthropic	Users can provide feedback from live sessions with models through these “thumbs up/down” mechanisms.
Thumbs Up/Thumbs Down function for feedback within Mistral’s Le Chat UI	Mistral	
Thumbs Up/Thumbs Down function for feedback within GPT’s UI	OpenAI	
Thumbs Up/Thumbs Down function for feedback within Gemini’s UI	Google	
Thumbs Up/Thumbs Down function for feedback within Command R’s UI	Cohere	

Anthropic's API Reference — Getting Help Section	Anthropic	Application developers and other users can contact the specific email to provide feedback or for other purposes.
Stability AI's Stable Diffusion Model Card — Contact Section	Stability AI	
Cohere's FAQ Page	Cohere	
Inflection Developer Dashboard	Inflection	
Support Downstream Dev of Azure AI to enable collection of aggregated metrics and user feedback for their deployment	Microsoft	Creating guides to support downstream level users to collect feedback and aggregated metrics for their own deployment.
4.2: Aggregating individual feedback records to have as summaries		
IBM watsonx Feedback Portal	IBM	Developers and other users can provide feedback through developer channels from a model hub's portal.
OpenAI Developer Forum	OpenAI	
Community discussion forum on Hugging Face for Phi-2	Microsoft	Developers and other users can provide feedback through public discussion posts and forums provided by a model hub/ hosting service like Hugging Face.
Community discussion forum on Hugging Face for StarCoder	BigCode	
Community discussion forum on Hugging Face for Mistral 7B	Mistral	
Community discussion forum on Hugging Face for Stable Diffusion	Stability AI	
Cohere's Community Discord Channel	Cohere	Users can utilize Cohere's public community forum for various engagements that may include feedback or support tickets.
4.3: Disclosing the process of following-up after going through a feedback process or redress mechanism		
Consent of Data Subjects of StarCoder 2's Governance Card	BigCode	Model Providers disclose what follow-up actions can be warranted post feedback submission.
Send Feedback with Gemini Apps	Google	
4.4: Create incentive structures to invite stakeholders to participate in the feedback process proactively		
Microsoft AI Bounty Program	Microsoft	"The Microsoft AI bounty program invites security researchers from across the globe to discover vulnerabilities in the new, innovative, Microsoft Copilot. Qualified submissions are eligible for bounty rewards from $2,000 to $15,000 USD. This bounty program is subject to these terms and those outlined in the Microsoft Bounty Terms and Conditions and our bounty Safe Harbor policy." "[Google] bug reports… assist our bug hunting community in effectively testing the safety and security of our AI products. Our scope aims to facilitate testing for traditional security vulnerabilities as well as risks specific to AI systems."
Google Bug Hunters: vulnerabilities in AI products	Google	

Endnotes

1 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Nicholas, Gabriel. “Grounding AI Policy: Towards Researcher Access to AI Usage Data.” *Center for Democracy and Technology*, 11 Sept. 2024, cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/.

Chan, Alan, et al. “Visibility into AI Agents.” *ArXiv.org*, 17 Apr. 2024, arxiv.org/abs/2401.13138.

Zhao, Wenting, et al. “WildChat: 1M ChatGPT Interaction Logs in the Wild.” *ArXiv.org*, 2024, arxiv.org/abs/2405.01470.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

Longpre, Shayne, et al. “The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources.” *ArXiv.org*, 2024, arxiv.org/abs/2406.16746.

Stein, Merlin, and Connor Dunlop. “Safe beyond Sale: Post-Deployment Monitoring of AI.” *Adalovelaceinstitute.org*, 2024, www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/.

2 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

3 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

Risto Uuk, et al. “Effective Mitigations for Systemic Risks from General-Purpose AI.” *ArXiv (Cornell University)*, 1 Jan. 2025, papers.ssrn.com/sol3/papers.cfm?abstract_id=5021463, https://doi.org/10.2139/ssrn.5021463.

4 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

5 Chmielinski, Kasia, et al. “The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers.” *Shorenstein Center*, 21 May 2024, shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/.

6 Winecoff, Amy A, and Miranda Bogen. “Improving Governance Outcomes through AI Documentation: Bridging Theory and Practice.” *ArXiv.org*, 13 Sept. 2024, arxiv.org/abs/2409.08960.

7 Bommasani, Rishi, et al. “A Path for Science- and Evidence-Based AI Policy.” *Understanding-Ai-Safety.org*, 2024, understanding-ai-safety.org/.

8 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Nicholas, Gabriel. “Grounding AI Policy: Towards Researcher Access to AI Usage Data.” *Center for Democracy and Technology*, 11 Sept. 2024, cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/.

Chan, Alan, et al. “Visibility into AI Agents.” *ArXiv.org*, 17 Apr. 2024, arxiv.org/abs/2401.13138.

Zhao, Wenting, et al. “WildChat: 1M ChatGPT Interaction Logs in the Wild.” *ArXiv.org*, 2024, arxiv.org/abs/2405.01470.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

Longpre, Shayne, et al. “The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources.” *ArXiv.org*, 2024, arxiv.org/abs/2406.16746.

Stein, Merlin, and Connor Dunlop. “Safe beyond Sale: Post-Deployment Monitoring of AI.” *Adalovelaceinstitute.org*, 2024, www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/.

9 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

10 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

Risto Uuk, et al. “Effective Mitigations for Systemic Risks from General-Purpose AI.” *ArXiv (Cornell University)*, 1 Jan. 2025, papers.ssrn.com/sol3/papers.cfm?abstract_id=5021463, https://doi.org/10.2139/ssrn.5021463.

11 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

12 Vipra, Jai, and Anton Korinek. “Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT.” *Brookings*, 7 Sept. 2023, www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt/.

13 “2023 March Audited Signed Accounts.” *Cifas.org.uk*, 15 Sept. 2023, www.cifas.org.uk/secure/contentPORT/uploads/documents/2023%20March%20Audited%20Signed%20Accounts.pdf.

14 “PAI’s Guidance for Safe Foundation Model Deployment.” *Partnership on AI*, 29 Oct. 2024, partnershiponai.org/modeldeployment/.

Nicholas, Gabriel. “Grounding AI Policy: Towards Researcher Access to AI Usage Data.” *Center for Democracy and Technology*, 11 Sept. 2024, cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/.

Chan, Alan, et al. “Visibility into AI Agents.” *ArXiv.org*, 17 Apr. 2024, arxiv.org/abs/2401.13138.

Zhao, Wenting, et al. “WildChat: 1M ChatGPT Interaction Logs in the Wild.” *ArXiv.org*, 2024, arxiv.org/abs/2405.01470.

Bommasani, Rishi, et al. “The Foundation Model Transparency Index.” *ArXiv.org*, 19 Oct. 2023, arxiv.org/abs/2310.12941.

Longpre, Shayne, et al. “The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources.” *ArXiv.org*, 2024, arxiv.org/abs/2406.16746.

Stein, Merlin, and Connor Dunlop. “Safe beyond Sale: Post-Deployment Monitoring of AI.” *Adalovelaceinstitute.org*, 2024, www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/.

15 McKernon, Elliot, et al. “AI Model Registries: A Foundational Tool for AI Governance.” *ArXiv.org*, 2024, arxiv.org/abs/2410.09645.

16 Howell, John, and Stephanie Ifayemi. “Policy Alignment on AI Transparency.” *Partnership on AI*, 9 Oct. 2024, partnershiponai.org/policy-alignment-on-ai-transparency/.

17 Srikumar, Madhulika, et al. “Risk Mitigation Strategies for the Open Foundation Model Value Chain.” *Partnership on AI*, 19 July 2024, partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/. See “Release Models with Digital Signatures or ‘Fingerprints’”

18 Srikumar, Madhulika, et al. “Risk Mitigation Strategies for the Open Foundation Model Value Chain.” *Partnership on AI*, 19 July 2024, partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/. See “Implement Disclosure Mechanisms for AI-generated Content.”

19 Srikumar, Madhulika, et al. “Risk Mitigation Strategies for the Open Foundation Model Value Chain.” *Partnership on AI*, 19 July 2024, partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/. See “Develop and Implement Durable Model-level Safeguards”

20 Srikumar, Madhulika, et al. “Risk Mitigation Strategies for the Open Foundation Model Value Chain.” *Partnership on AI*, 19 July 2024, partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/. See “Monitor Misuses, Unintended Uses, and User Feedback”

21 Howell, John, and Stephanie Ifayemi. “Policy Alignment on AI Transparency.” *Partnership on AI*, 9 Oct. 2024, partnershiponai.org/policy-alignment-on-ai-transparency/.

22 Chmielinski, Kasia, et al. “The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers.” *Shorenstein Center*, 21 May 2024, shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/.

23 Paeth, Kevin, et al. “Lessons for Editors of AI Incidents from the AI Incident Database.” *ArXiv.org*, 2024, arxiv.org/abs/2409.16425.