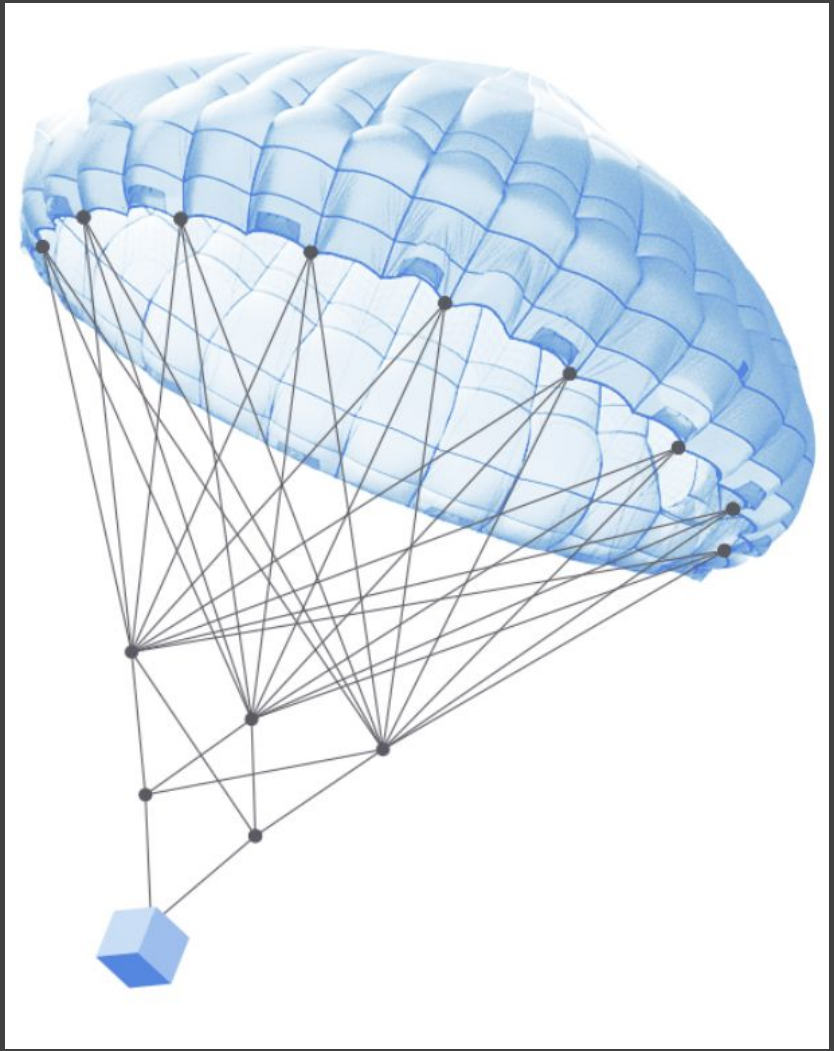
The Alan Turing Institute

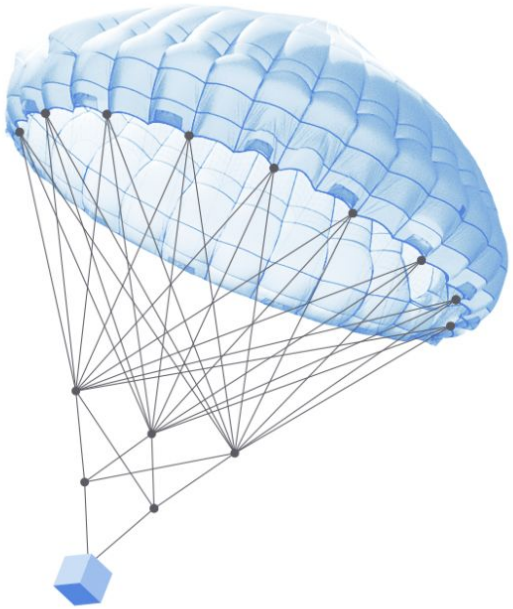
Best Practice for Responsible Foundation Models

What Should Developers Do and How You Can Help

Dr Carolyn Ashurst



PARTNERSHIP ON AI



PAI's Guidance for Safe Foundation Model Deployment

A Framework for Collective Action

Partnership on AI's (PAI) Guidance for Safe Foundation Model Deployment is a framework for model providers to responsibly develop and deploy a range of AI models, promote safety for society, and adapt to evolving capabilities and uses.

OPEN FOR PUBLIC COMMENT

SUBMIT FEEDBACK

Agenda

1. Developing the PAI Guidance for Safe Foundation Model Deployment
2. Deployment Guidance at a glance
3. Discussion

What are the PAI Guidelines and what is their purpose?

To provide guidance to foundation model developers:

- Based on the capabilities of the model
- Based on how the model will be released
- For different stages of the development/deployment cycle
- That can be updated over time as a living document

Why do we need guidelines for foundation models?

- There are many potential risks and harms from foundation models and their downstream applications
- Increasing developments and deployments means impacts will likely accelerate over time
- Best practice and regulation are struggling to keep up

Development timeline

- 2022/2023: Safety steering committee meetings
- April 2023: IBM-PAI workshop
- June 2023: Working group formed
- June 2023: Workshop on wider governance
- October 2023: Draft guidance released for public comment
- January 2023: Deadline for public comments
- 2024: ~~Final~~ Next version launched

Working Group

Andrew Strait
Ada Lovelace Institute

Markus Anderljung
GovAI

Joslyn Barnhart
DeepMind

Jerremy Holland
Director of AI Research
Apple

Carolyn Ashurst
The Alan Turing Institute

Harrison Rudolph
Meta

Jared Mueller
Anthropic

Joshua New
IBM

Jessica Young
Microsoft

Anthony Barrett
Berkeley CLTC

Yolanda Lannquist
The Future Society

Ellie Evans
Cohere

Reena Jana
Google

Main audiences

Industry Champions	The guidelines **empower key champions and decision-makers within industry and other model provider organizations** with a collectively agreed upon framework to guide deployment decisions of large-scale AI.
Policymakers	The guidelines aim to **complement and inform** other individual and collaborative efforts, **including policy and efforts like the Frontier Model Forum**, to develop enforceable guidance or tooling in alignment with the protocols.

MODEL PROVIDERS

The Model Deployment Guidance distinguishes model providers from actors in the broader AI ecosystem (seen below) as those training foundational models that others may build on.

ECOSYSTEM ACTOR	ROLE DESCRIPTION
Compute / Hardware Providers	Providing underlying compute power to train and run models
Cloud Providers	Providing underlying cloud infrastructure to support training of and deployed models
Data Providers	Providing training datasets (intentionally or unintentionally) for model providers, may also be model providers
Model Providers	**Training foundational models (proprietary or open-source) that others may build on**
Application Developers (or: Service Developers, Model Integrators)	Building applications and services on top of foundational models
Consumers and/or Affected Users	Consumers (B2C) who are end-users of services built on top of foundational models Affected Users may be impacted or implicated in the use of AI (e.g. medical AI Consumers are doctors, and Affected Users are patients)

User Flow

1 → 2 → 3 → ★

What kind of model?

What kind of release?

Update or new model?

Generated Guidance

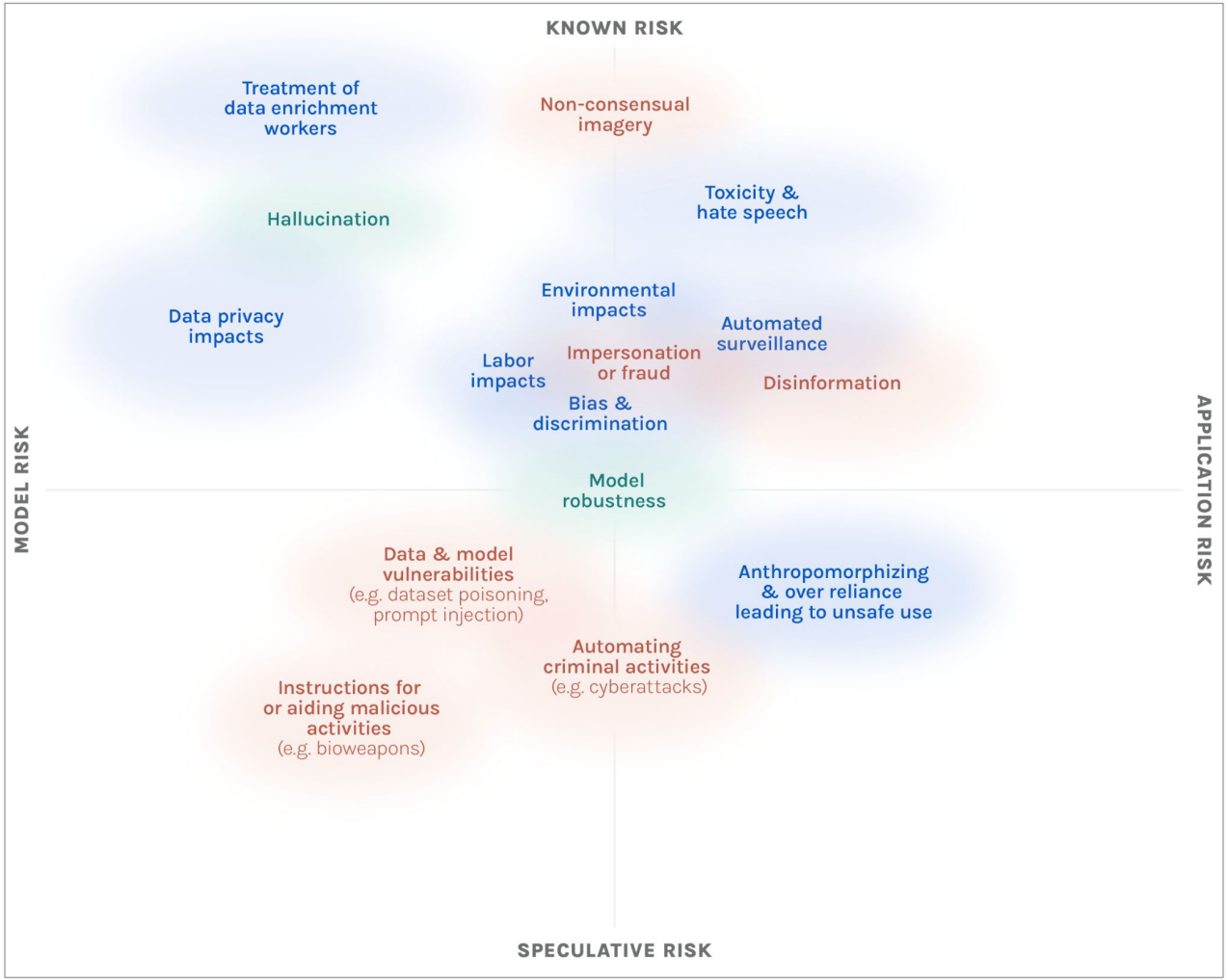
Each protocol:

- Baseline practices
- Recommended practices
- Implementation resources

Protocols in four sections:

Research & Development • Pre-Deployment • Post-Deployment • Societal Impact

Risk Landscape



SUB-CATEGORIES OF RISKS

- **Malicious uses:** Risks of intentional misuse or weaponization of models to cause harm
- **Societal risks:** Potential harms that negatively impact society, communities and groups
- **Other Risks:** Risks distinct from the above categories

Framework for tailored guidelines

There are a total of 22 guidelines. Not all model types and releases are treated equally within the Guidance paradigm. The guidelines are more extensive for **more capable** models and **more available** release types.

	A. Open Release	B. Restricted Release	C. Research Release	D. Closed Dev
1. Specialized Narrow Purpose				
2. Advanced Narrow & GPAI				
3. Paradigm-shifting Frontier				

HIGHER BURDEN

3A is considered too risky to be included

TYPE OF FOUNDATION MODEL	EXAMPLES
1. Specialized Narrow Purpose Models designed for narrowly defined tasks or purposes with limited general capabilities for which there is lower potential for harm across contexts.	• Music generation models

TYPE OF FOUNDATION MODEL	EXAMPLES
1. Specialized Narrow Purpose Models designed for narrowly defined tasks or purposes with limited general capabilities for which there is lower potential for harm across contexts.	• Music generation models
2. Advanced Narrow and General Purpose Models with generative capabilities for synthetic content like text, image, audio, video. Can be narrow purpose focused on specific tasks or modalities or general purpose. Also covers **some narrow purpose models focused on scientific, biological or other high consequence domains.** Encompasses general purpose models capable across diverse contexts, like chatbots and multimodal models.	• Text to speech generation (Meta Voicebox) • Text to video generation • Code generation (Copilot etc) • LLaMA-2; LLaMA Chat • Galactica • ChatGPT • GPT4 • BLOOM • GPT

TYPE OF FOUNDATION MODEL	EXAMPLES
1. Specialized Narrow Purpose Models designed for narrowly defined tasks or purposes with limited general capabilities for which there is lower potential for harm across contexts.	• Music generation models
2. Advanced Narrow and General Purpose Models with generative capabilities for synthetic content like text, image, audio, video. Can be narrow purpose focused on specific tasks or modalities or general purpose. Also covers **some narrow purpose models focused on scientific, biological or other high consequence domains.** Encompasses general purpose models capable across diverse contexts, like chatbots and multimodal models.	• Text to speech generation (Meta Voicebox) • Text to video generation • Code generation (Copilot etc) • LLaMA-2; LLaMA Chat • Galactica • ChatGPT • GPT4 • BLOOM • GPT
3. Paradigm-shifting or Frontier models Cutting edge models that significantly advance capabilities across modalities compared to the current state of the art.	• Extremely large multimodal models • Text to action

TYPE OF FM RELEASE	ILLUSTRATIVE EXAMPLES
1. Open Access Models released publicly with full access to components including weights, code, and data. Can be free or commercially licensed. Access can be downloadable or via other cloud, hosted or MLOps providers.	• GPT-J (free downloadable access) • Llama 1 (free available by request) • Llama 2 (commercial license download)

TYPE OF FM RELEASE	ILLUSTRATIVE EXAMPLES
1. Open Access Models released publicly with full access to components including weights, code, and data. Can be free or commercially licensed. Access can be downloadable or via other cloud, hosted or MLOps providers.	• GPT-J (free downloadable access) • Llama 1 (free available by request) • Llama 2 (commercial license download)
2. Restricted API and Hosted Access Models available only through a controlled API, cloud platform, or hosted through a proprietary interface, limiting access. Can include gradual staged releases. Does not provide direct possession of the model. Both allow restricting access and monitoring usage to mitigate risks.	• GPT-2 (staged release) • GPT-3 (API) • DALL-E 2 (API) • Midjourney (Hosted) • ChatGPT (Hosted)

TYPE OF FM RELEASE	ILLUSTRATIVE EXAMPLES
1. Open Access Models released publicly with full access to components including weights, code, and data. Can be free or commercially licensed. Access can be downloadable or via other cloud, hosted or MLOps providers.	• GPT-J (free downloadable access) • Llama 1 (free available by request) • Llama 2 (commercial license download)
2. Restricted API and Hosted Access Models available only through a controlled API, cloud platform, or hosted through a proprietary interface, limiting access. Can include gradual staged releases. Does not provide direct possession of the model. Both allow restricting access and monitoring usage to mitigate risks.	• GPT-2 (staged release) • GPT-3 (API) • DALL-E 2 (API) • Midjourney (Hosted) • ChatGPT (Hosted)
3. Closed Development Models developed confidentially within an organization first, with highly limited releases for internal evaluation or restricted external testing before any potential public availability.	

TYPE OF FM RELEASE	ILLUSTRATIVE EXAMPLES
1. Open Access Models released publicly with full access to components including weights, code, and data. Can be free or commercially licensed. Access can be downloadable or via other cloud, hosted or MLOps providers.	• GPT-J (free downloadable access) • Llama 1 (free available by request) • Llama 2 (commercial license download)
2. Restricted API and Hosted Access Models available only through a controlled API, cloud platform, or hosted through a proprietary interface, limiting access. Can include gradual staged releases. Does not provide direct possession of the model. Both allow restricting access and monitoring usage to mitigate risks.	• GPT-2 (staged release) • GPT-3 (API) • DALL-E 2 (API) • Midjourney (Hosted) • ChatGPT (Hosted)
3. Closed Development Models developed confidentially within an organization first, with highly limited releases for internal evaluation or restricted external testing before any potential public availability.	
4. Research Release Models released in a restricted manner to demonstrate research concepts, techniques, components such as publishing ingredients, demos, fine-tuned versions of existing models, and incremental progress.	• Stanford Alpaca • Megatron-Turing NLG • Jurassic-1 - AI21 Labs • GLIDE • ActGPT?

“Significant Update” Selection

Models that continue major development post-deployment by significantly expanding capabilities, necessitating renewed governance.

Features:

- Updates drastically enhance model architecture, knowledge scope, or modalities.
- Brings capabilities meaningfully beyond the initial release.
- Poses novel risks requiring additional evaluation.
- Meaningfully expands potential beneficial and harmful applications.

Examples:

- Switching to a drastically larger model architecture.
- Adding entirely new modalities like text+video after initial text-only.
- Enabling connections to live databases that greatly expand knowledge scope.

Guidelines: Frontier x Restricted Release

Research & Development

- Scan for novel risks
- Practice responsible iteration
- Assess upstream security vulnerabilities
- Publish a “Pre-Systems Card”
- Establish risk management structures & processes

Pre-Deployment

- Internally evaluate models for safety
- Conduct external model evaluations for safety
- Undertake red teaming and share findings
- Publicly report model impacts and “key ingredients”
- Provide downstream use documentation
- Establish safeguards to restrict unsafe use

Post-Deployment

- Monitor deployed systems
- Implement incident reporting
- Establish decommissioning policies
- Develop transparency reporting standards

Societal Impact

- Support third party inspection of models and training data
- Responsibly source all labour including data enrichment
- Conduct human rights due diligence
- Establish feedback mechanisms across the AI value chain
- Disclose synthetic content
- Measure & disclose environmental impacts
- Measure & disclose severe labor market risks

Guidelines: Specialized x Open Release

Research & Development

- Scan for novel risks
- Assess upstream security vulnerabilities
- Establish risk management structures & processes

Pre-Deployment

- Internally evaluate models for safety
- Publicly report model impacts and “key ingredients”
- Provide downstream use documentation
- Establish safeguards to restrict unsafe use

Post-Deployment

- Monitor deployed systems

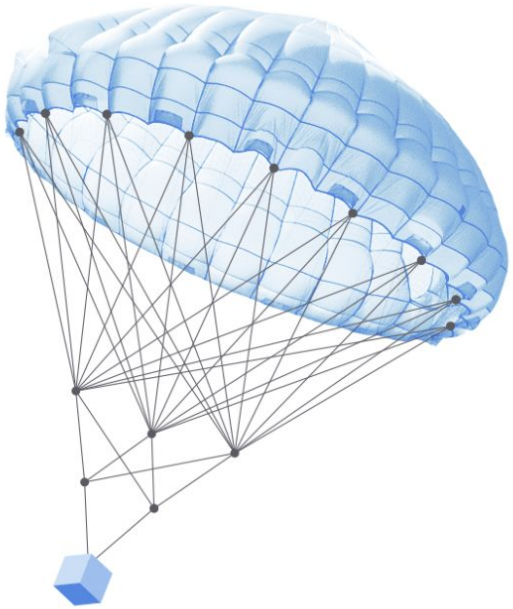
Societal Impact

- Responsibly source all labour including data enrichment
- Establish feedback mechanisms across the AI value chain

partnershiponai.org/modeldeployment

Search for: pai foundation guidance

PARTNERSHIP ON AI



PAI's Guidance for Safe Foundation Model Deployment

A Framework for Collective Action

Partnership on AI's (PAI) Guidance for Safe Foundation Model Deployment is a framework for model providers to responsibly develop and deploy a range of AI models, promote safety for society, and adapt to evolving capabilities and uses.

OPEN FOR PUBLIC COMMENT

SUBMIT FEEDBACK

Generate Custom Guidance

This guidance assists foundation model providers: organizations developing AI systems trained on broad datasets to power a wide variety of downstream uses and interactive interfaces.

1 Choose your foundation model

In cases where models do not fit into one category, choose the model type of higher capability.

Specialized Narrow Purpose +

Advanced Narrow and General Purpose +

Paradigm-shifting or Frontier +

My foundation model is: Select Model Type

2 Choose your type of release

Choose the intended initial release method. For phased rollouts, select the current stage and revisit this guidance as release plans progress.

Open Access +

Restricted API and Hosted Access +

Closed Development +

Research Release +

My release type is: Select Release Type

3 Is it an update?

Significantly Enhanced Update Launch +

If the release is a significant update to an existing model, you are encouraged to renew governance processes as needed per the guidance for your model and release type.

☐ Yes, it is an update

Frontier & Restricted Foundation Models

These guidelines for model providers focus on base models and their interactive interfaces. Further risk evaluations — that address specific use cases and domains — by downstream application developers are still important.

Research & Development

- Scan for novel or emerging risks +
- Practice responsible iteration +
- Assess upstream security vulnerabilities +
- Produce a "Pre-Systems Card" +
- Establish risk management and responsible AI structures for foundation models +

Pre-Deployment

- Internally evaluate models for safety +
- Conduct external model evaluations to assess safety +
- Undertake red-teaming and share findings +
- Publicly report model impacts and "key ingredient list" +
- Provide downstream use documentation +
- Establish safeguards to restrict unsafe uses +

Post-Deployment

- Monitor deployed systems +
- Implement incident reporting +
- Establish decommissioning policies +
- Develop transparency reporting standards +

Societal Impact

- Support third party inspection of models and training data +
- Responsibly source all labor including data enrichment +
- Conduct human rights due diligence +
- Enable feedback mechanisms across the AI value chain +
- Measure and disclose environmental impacts +
- Disclose synthetic content +
- Measure and disclose anticipated severe labor market risks +

SUBMIT FEEDBACK

Pre-Deployment

Internally evaluate models for safety +

Conduct external model evaluations to assess safety +

Undertake red-teaming and share findings -

DESCRIPTION

Implement red teaming that probes frontier models for potential malicious uses, societal risks and other identified risks prior to release. Address risks and responsibly disclose findings to advance collective knowledge.

BASELINE PRACTICES i

- Perform internal and external red teaming across model capabilities, use cases, and potential harms including dual-use risks using techniques such as adversarial testing, vulnerability scanning, and surfacing edge cases and failure modes.
- Conduct iterative red teaming throughout model development. Continuously evaluate results to identify areas for risk mitigation and improvements, including for planned safeguards.
- Commission external red teaming by independent experts such as domain experts and affected users to surface gaps. Select external red teamers to incentivize the objective discovery of flaws and ensure adequate independence.
- Address identified risks and adapt deployment plans accordingly based on learnings from pre-deployment evaluations.
- Responsibly disclose findings, aligned with guidance below on public reporting.

RECOMMENDED PRACTICES i

- Collaborate across industry, civil society, and academia to advance red teaming methodologies and responsible disclosures.

Submit Feedback

We invite you to actively participate in shaping the future of responsible development of foundation models. Your feedback is essential to enhance our Guidance for Safe Foundation Model Deployment collectively. By sharing your insights, you contribute to a more robust and inclusive framework.

We're actively seeking public comments on the Model Deployment Guidance until January 15, 2024, with plans to release an updated version later in the year.

To learn about future opportunities to engage with PAI's Guidance for Safe Foundation Model Deployment, including virtual convenings where you can provide direct feedback on the guidelines and partner with us on our next steps, please complete the form. Your engagement is vital to realizing our shared goal of ensuring the safe and responsible deployment of foundation models.

First Name *

Last Name *

Email *

Affiliation *

Please write the name of your organization in full

Title

Comments

Share your feedback and/or questions on the Model Deployment Guidance, including the framework for model capabilities and release types. If you're interested in collaborating, please describe your organization and your interest

☐ **Join mailing list to hear about virtual convenings and more**

Submit

Next steps

- October 2023: Draft guidance released for public comment
- January 2023: Deadline for public comments
- 2024: ~~Final~~ Next version launched
- 2024+ Supporting applying the framework in practice
- 2024+ Exploring operationalization of the framework
- 2024+ Exploring responsibility across the value chain

Challenges

1. What can/can't voluntary guidelines achieve?
2. How should responsibility fall within the development pipeline?
3. What should best practice look like for general purpose models?
4. What can we do in response to the regulatory pacing problem?
5. What can we do in response to developments outpacing best practice?

Thanks!

Slides credit: Thanks to PAI for the initial slide deck from which this deck was adapted