PARTNERSHIP ON AI

Policy Alignment on AI Transparency

Analyzing Interoperability of Documentation Requirements across Eight Frameworks

John Howell
Stephanie Ifayemi

Contents

Introduction

As the world grapples with harnessing the benefits of increasingly powerful foundation models and managing their attendant risks, we are seeing an increasing pace of policy development in the pursuit of these goals. This is happening at the national, regional, and international levels and ranges from high-level statements of principle, such as the OECD Recommendation of the Council on Artificial Intelligence and the UN Global Digital Compact[A]; through non-binding standards/frameworks, such as the NIST Artificial Intelligence Risk Management Framework ("NIST AI RMF") and its Generative AI Profile; to binding requirements under the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence ("AI Executive Order"), the EU Artificial Intelligence Act ("EU AI Act"), and the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law ("COE AI Convention").

These existing frameworks continue to be built out through the development of guidance materials,[B] codes of practice,[C] standards,[D] and monitoring mechanisms.[E] New frameworks continue to be developed or are under consideration in an increasing number of countries. This policy action is both necessary and timely. However, without coordinated efforts, there is a risk it will lead to an incoherent patchwork of frameworks, which build fragmented understandings of good practice. This paper explores the following questions: Are current policy frameworks for foundation model documentation interoperable? What challenges to interoperability can we see on the horizon, and what impact might this have on best practice and accountability across borders? How can interoperability between policies and laws governing foundation models be promoted as these frameworks continue to develop? More specifically, when we consider documentation requirements, to what extent do they align, and what are the risks if they do not?

The term "interoperability" can be used in a number of contexts. It can refer to:

- **Policy interoperability:** This concerns how similar or well-aligned the provisions of policy frameworks are. One key measure that is often used is to focus on how easy it is to comply with multiple frameworks and whether compliance with one framework will make it easier to comply with others. Another, more important, way to view interoperability for Partnership on AI ("PAI") is to focus on creating regulatory consistency. This is to converge on and establish accountability around good practice across countries while facilitating the ability of stakeholders (e.g., auditors, civil society) to compare, study, and evaluate models across borders.
- **Institutional interoperability:** This concerns how aligned the functions or operations of institutions are. A high degree of alignment can allow institutional cooperation or mutual recognition of each other's functions. This can supplement policy interoperability.
- **Technical interoperability:** This concerns how well technologies, systems, or products work with each other.

[A] The final text of the Global Digital Compact was agreed at the UN Summit of the Future in September 2024.

[B] E.g., the AI Executive Order mandates the development of a range of guidance documents such as the recently released NIST Generative AI Profile to the NIST AI RMF.

[C] E.g., the EU AI Act mandates the development of harmonized standards for both high-risk AI systems and general-purpose AI models.

[D] The EU AI Act provides for the development of Codes of Practice for General Purpose AI models.

[E] E.g., the OECD is developing a monitoring mechanism for the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.

This report concerns policy interoperability (and to a degree, how institutional interoperability can complement it).

Why does PAI care about interoperability?

Interoperability should not be an end goal in and of itself. Rather, it holds the potential to promote a number of other objectives, including the beneficial development and use of AI if a good set of requirements is used as the basis for achieving interoperability. PAI's core interest in this work is the potential for interoperability efforts to promote accountability and best practices across borders—particularly for documentation, which is a critical tool for accountability. For example, consistent documentation practices aligned on a good set of requirements support an international auditing ecosystem, which could provide transparency about model risks across borders. Promoting these efforts aligns with PAI's theory of change by using our multistakeholder approach to inform policy innovation and ultimately change industry practice by ensuring we set a good benchmark for foundation model governance across countries.

Aligning on good practices for documentation is particularly important at this point in time as foundation models continue to increase in capability and are deployed in new applications, such as AI agents, which will increase the need for policies promoting accountability across borders. We want to (i) firstly inform the development of good documentation practices, building on PAI's resources and research, and then (ii) drive forward policy and institutional interoperability, fostered by a strong baseline of good practices, where governments converge around what 'good' looks like, and hold organizations accountable.

What do we mean by good documentation practices?

This work builds on PAI's research to establish best practices, including for documentation through our ABOUT ML work. By setting a good baseline for practice, there is an opportunity to facilitate accountability and potentially protect people across borders while still achieving various economic and innovation benefits. In other words, (a) incorporating good documentation practices in policy frameworks promotes accountability and other benefits, and (b) interoperability efforts should be directed towards aligning 'good' documentation requirements across national, regional, and international frameworks so these best practices are shared.[F]

Avoiding a lowest common denominator is critical

It is important that efforts to promote interoperability are not used to promote agreement on a lowest common denominator for documentation practices but rather to set and align around best practices. Committed efforts from industry as well as national governments, civil society, and other stakeholders will be necessary to achieve this. While there will be some legitimate differences between policy approaches in different countries, that does not mean a good baseline cannot and should not be set.

[F] PAI's Guidance for Safe Foundation Model Deployment contains recommendations for best practices for foundation model documentation. While there is not significant detail about the best form or content of documentation artifacts in current policy frameworks, the literature describes a number of artifacts, some of which (such as model cards and datasheets) have seen significant rates of adoption and become "quasi-standard"; though the level of detail included in them in practice varies, (see e.g., National Telecommunications and Information Administration, Artificial Intelligence – Accountability Policy Report, March 2024, p. 30.)

What this report is about

This report examines current and potential near-term interoperability challenges between a number of leading policy frameworks for foundation models.

It focuses on documentation requirements in those frameworks. More information about how the in-scope frameworks were chosen and the reasons for our focus on documentation are given below. By documentation, we are specifically referring to information that is recorded for an external audience (including regulators, downstream developers, and the wider public, though different documentation may have different intended audiences), and to specific documentation artifacts such as model cards and datasheets. These forms of documentation are critical as they provide information about foundation models—including their "key ingredients" and what testing and evaluations they have been subject to—which is important to achieve accountability and transparency within and across jurisdictions.

This report includes a review of select leading international policy frameworks aimed at addressing foundation model risks.

It maps and compares the documentation requirements in those frameworks, considers the next steps being taken to add more detail under those frameworks, assesses what interoperability issues exist or are foreseeable, and considers what steps might promote interoperability, best practice, and accountability moving forward. This report also considers the role of specialized institutions set up (and proposed) in the countries under review—notably AI Safety Institutes ("AISIs") and the EU AI Office—and considers what role they might play in promoting interoperability and establishing best practices.

Interoperability has the potential to bring many benefits if done well and established on the basis of multistakeholder input. In particular, interoperable policy frameworks can promote accountability across borders for all actors through the AI value chain. This is essential to allow the development of a global AI ecosystem, to build trust in foundation models and products built on them, and ultimately to address risk and protect fundamental rights. Interoperability efforts are an opportunity to promote best practices in the pursuit of the shared goals of promoting the beneficial development of foundation models. Alignment of documentation requirements is particularly important in this regard (though it may be that not all other aspects of AI governance require global alignment to the same degree).

Interoperable policy frameworks can promote accountability across borders for all actors through the AI value chain.

Background and methodology

The work plan leading to this report was developed with guidance from PAI's Policy Steering Committee, composed of global AI leaders and experts. This report has been informed through desk research and consultations with experts from industry, civil society, academia, and non-profit organizations, drawn from PAI's partner and wider stakeholder networks. We tested our initial thinking in a virtual multistakeholder workshop in August 2024. This report would not have been possible without the support and diverse expertise of these stakeholders, including those referenced at the end of the report. The views and recommendations in this report remain those of PAI.

Key findings, recommendations, and questions for the future

Summary of findings

The key findings of this report are:

1. **Interoperability and collaboration are explicitly included as policy goals in a number of these international frameworks** (e.g., the EU AI Act, the UK government's "Pro-Innovation Approach to AI Regulation," and the G7 Hiroshima AI Code of Conduct). That is welcome but will require concerted, ongoing efforts to be realized, as there is no current agreement on how to put this high-level goal into practice.
2. **These frameworks also emphasize the importance of documentation in achieving key policy objectives; however, they remain light on details** about the form and content of documentation to be produced for foundation models. This means there are not yet significant interoperability issues for documentation in the US, EU, UK, and multilateral frameworks (such as the Hiroshima Code of Conduct) reviewed in this report. However, the policy landscape is evolving rapidly, and initiatives are underway to develop more detailed requirements (the most noteworthy being under the EU AI Act). That means that areas of inconsistency could arise as more detailed rules and guidance are introduced. We see a particular risk and opportunity on the horizon with the development of more detailed requirements under the EU AI Act and potentially through the G7's Hiroshima AI Process.
3. **There are a number of steps that could be taken to advance interoperability now and in the future**, such as leveraging existing and proposed forums, mechanisms, and processes. These include promoting collaborative research initiatives, supporting the engagement of diverse stakeholders in policy initiatives, and working to promote alignment between ongoing policy initiatives. These proposed steps are discussed in more detail later in this report.
4. **An early focus should be establishing agreed capability thresholds for regulation**, as well as providing international consistency about which foundation models are captured by regulatory/policy frameworks to underpin efforts to align requirements in those frameworks, including for documentation.
5. **While there are some challenges to relying on international standardization processes to align AI policy frameworks, they remain an important plank in that effort.** Nations (and regional entities such as the EU) should commit to developing

and adopting international standards to address foundation model risks to the extent possible.

6. **Harmonizing key documentation requirements across national, regional, and international foundation model policy frameworks**—and, in particular, harmonizing the form and content of documentation artifacts—should be made a priority in standardization and other interoperability efforts. Detailed guidance for dataset documentation should be a particular priority. This work does not have to start from scratch and can build on artifacts described in the literature, including model cards and datasheets.

7. **The lack of consensus on the best approaches to manage AI risks is a significant challenge to developing interoperable frameworks**, including for documentation. Collaboration in the science of AI safety will promote interoperability efforts. Bodies and initiatives such as the current and announced AI Safety Institutes, the Network of AI Safety Institutes, the Interim International Scientific Report on the Safety of Advanced AI, and the recently announced UN International Scientific Panel on AI are all promising avenues for taking this work forward. However, efforts will be needed to ensure the work of these initiatives is aligned to prevent further policy fragmentation by implementing measures to act on their findings and outputs.

8. **The existing and newly announced AI Safety Institutes have significant potential to provide a foundation for agreement on AI safety** through research, the development and conduct of evaluations, advancing the science of AI Safety, and creating common approaches to safety and documentation through collaboration and information sharing. A shared understanding of risks and how to address them can underpin common policy goals and, subsequently, alignment in the development of more detailed policy frameworks. Governments should ensure the Safety Institutes and similar entities that may be established have the necessary mandates and resources to fulfill these functions.[G] Model providers should make every effort, to the extent practicable, to cooperate with AI Safety Institutes and participate in their processes. Efforts will also be needed to ensure the work of these bodies does not compete or conflict with work being undertaken by other institutions such as the OECD.

9. **Participation by civil society and the global community is needed in all major foundation model policy initiatives** if we are to ensure that they lead to alignment around best practices and that the agenda for global interoperability is not set by a comparatively small group of nations from the Global North. Leading national and regional frameworks are likely to influence the direction of global policy. Incorporating a wide range of perspectives and expertise in developing these leading frameworks will promote good and interoperable policy development. Governments, international standards development organizations ("SDOs"), industry and other actors should support engagement by these stakeholders. The France AI Action Summit will be a good forum to advance this objective, including through its Public Interest AI and Global AI Governance tracks.

[G] The "similar entities" referred to here include the EU AI Office, which has functions that significantly overlap with existing AI Safety Institutes (in the case of the EU, these functions are in addition to a wide range of other functions and powers).

Recommendations

More discussion of these recommendations can be found in the recommendations section at the end of this report.

1. **National governments and the EU should prioritize cooperation in setting thresholds** for identifying which foundation models require additional governance measures, including through supporting the OECD's work on this issue. The AI Summit Series could also be used to take this forward. Agreement on thresholds is important for promoting interoperability for documentation (and other) requirements. The safety measures required to manage risks posed by powerful models are likely to depend on the capabilities of the in-scope models and the categories of risk that the frameworks are intended to address. Agreeing on a common definition and thresholds for the models covered by policy frameworks should flow through to greater alignment between the frameworks, including in relation to documentation requirements.

2. A. **The G7 Presidency should continue developing the Hiroshima Code of Conduct into a more detailed framework** to provide more detail about thresholds, relevant risks, and the form and content of documentation artifacts. This work should be a focus of Canada's G7 Presidency in 2025, including aligning closely with the EU Codes of Practice development timeline. In doing this, it should seek input from foundation model providers, civil society, academia, and other stakeholder groups equally. This will strengthen accountability by making the proposed monitoring mechanism—currently being developed by the OECD—more robust. It could also be a mechanism to promote convergence between the Code of Conduct and pending Codes of Practice under the EU AI Act. That convergence could be a powerful lever for wider interoperability.

   B. In developing and approving the initial Codes of Practice under the EU AI Act, participants in the development process, **the AI Office, the AI Board, and the EU Commission should adopt interoperability with other leading frameworks as a key objective**, to the extent practicable. This would mean other frameworks and the documentation artifacts they call for, such as technical documentation under the G7 Hiroshima Code of Conduct, are considered when the Codes of Practice are being formulated. The involvement of non-EU model providers, experts, and civil society organizations will help advance this objective. Steps towards this objective will depend on the pace of development of various frameworks.

3. A. To support the development of standardized documentation artifacts such as dataset documentation and technical documentation, **Standards Development Organizations should ensure that their processes are informed by appropriate sociotechnical expertise, diverse perspectives, as well as required resources.** To that end, SDOs, industry, governments, and other bodies should invest in capacity building for civil society and academic stakeholders to engage in standards-making processes, including to ensure participation from the Global South. That could include engaging in more active outreach and providing financial and logistical support. This is critical to ensure multistakeholder, sociotechnical, and global expertise informs these processes. Governments should consider mirroring initiatives such as the UK's AI Standards Hub to achieve this goal.

B. **The development of standardized documentation artifacts for foundation models, such as datasheets, should be a priority in AI standardization efforts.** This would promote internationally comparable documentation requirements for foundation models—promoting interoperability and establishing a baseline for best practice internationally. Given the bottom-up process in standards-making, which is currently largely led by individuals and industry proposing work items, there is a significant need for industry to prioritize and support efforts to develop standards addressing this topic. As with all standardization processes, it will also be important to support meaningful engagement by non-industry and non-Global North stakeholders.

4. **International collaboration and research initiatives should prioritize research that will support the development of standards for foundation model documentation artifacts**, including dataset documentation and technical documentation. Documentation is a key feature of foundation model policy requirements, and common requirements for artifacts will directly improve interoperability. It will also make comparisons between models from different countries easier, promoting accountability and innovation.

5. A. **National governments should continue to prioritize both international dialogue and collaboration on the science of AI Safety** through initiatives such as the AI Summit series, the Interim International Scientific Report on the Safety of Advanced AI, and the UN International Scientific Panel on AI, however with more specificity and tracking of progress on commitments that will foster good practice. A key priority in preparations for the AI Action Summit in February 2025 should be ensuring there is a pathway to take this work forward following the release of the finalized International Report at that event. Documentation needs now and in the future provide a strong starting point for future work, including to deliver on commitments related to interoperability, accountability and transparency. This work will inform a common understanding of what should be included in documentation artifacts to promote accountability and address foundation model risks. Aligning the work of the AI Summits (including the International Scientific Report) and the UN International Scientific Panel on AI, when established, to ensure the responsibilities and scope of work for each are clear will help ensure they fulfill their potential as drivers for consensus and best practice. Fragmentation of collaborative research initiatives, or divergence of the work of these bodies, could contribute to divergence in policy implementation approaches.

B. **National governments should support the creation/development of AI Safety Institutes (or equivalent bodies) and ensure they have the resources, functions, and powers necessary to fulfill their core tasks** (particularly advancing the science of evaluation as a first focus). Efforts should be made to align the functions of these bodies with those common among existing AI Safety Institutes. This will promote efforts to develop trusted mechanisms to evaluate advanced foundation models and may, at a later stage, lead to the potential to work towards "institutional interoperability," for example through mutual recognition of evaluations.

C. **The Network of AISIs (and bodies with equivalent or overlapping functions, such as the EU AI Office) should be supported, and efforts should be made to expand its membership.** Consideration should be given to how the Network could support broader AI safety research initiatives—for instance, through sharing expertise gained by constituent AISIs in performing their functions and inputting to other initiatives such as the newly announced UN International Scientific Panel on AI.

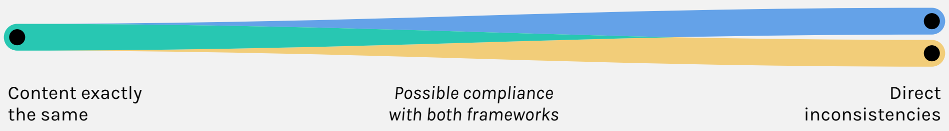
Interoperability and why it matters

What do we mean by "interoperable"?

This report considers the **interoperability of policy frameworks**, including legal and regulatory frameworks. It is not directly concerned with technical interoperability between AI systems (though that could be one aim of interoperable policies).

"Interoperability" is not a term of art. In this paper, we use it as a general term to refer to the level of compatibility or consistency between policy frameworks. This can be considered from the perspective of how similar the requirements of various frameworks are or how easy it is to apply or comply with multiple frameworks.

Spectrum of Interoperability
Consistency between frameworks



With this in mind, **interoperability is not all or nothing.** Rather, it exists on a spectrum. At one end of the spectrum, the content or requirements of two frameworks might be exactly the same. At the other extreme, two frameworks might contain direct inconsistencies, making it impossible to follow both at the same time. Or consistency between frameworks can lie somewhere in between these extremes—if compliance with two frameworks involves following and documenting two different sets of requirements, or completing two different sets of evaluations or certification processes, compliance with both may be possible, but will be significantly more difficult and lead to inconsistent practices in different places. In practice, it is this kind of interoperability challenge which is most likely to arise. At the same time, it is not reasonable to aim for identical regulation across jurisdictions. There are legitimate differences between regulatory approaches in different countries.[H] Interoperability efforts should be directed towards aligning the substance of policy and regulatory frameworks.

International interoperability initiatives are distinct from, though related to, international governance initiatives. There is a growing literature[I] discussing whether there is

[H] See e.g., The Brookings Institution's Strengthening International Cooperation on AI

[I] E.g., International AI Institutions: A Literature Review of Models, Examples, and Proposals; "International Institutions for Advanced AI"; Microsoft's Global Governance: Goals and Lessons for AI a need for global AI governance, including new institutions to set, monitor, and/or enforce global norms (such as an "IPCC for AI"[J]). A focus on interoperability starts with current and proposed national and multilateral frameworks and seeks ways to promote compatibility between them. This work can involve multilateral forums but does not assume the endpoint will necessarily be an internationally integrated hierarchy of governance institutions. Rather, the goal is that the different mechanisms designed by various institutions individually align and do not conflict.

Current frameworks set out **norms** in the form of laws, rules, or recommendations for model providers (and others in the AI value chain). They also establish **institutions** and **oversight mechanisms**. These institutions can play various roles, including developing subsidiary regulations or guidelines, maintaining registers of certain models, evaluating models and monitoring providers for safety or compliance with legal or non-binding norms, and requiring providers to implement mitigations. Interoperability can involve aligning either or both the **underlying frameworks** and the **functions and activities** of oversight bodies.

Why does interoperability matter?

PAI's consultations, as reflected in this report, revealed widespread agreement about the importance of policy interoperability. This support was shared across stakeholder groups (industry, civil society, and academia). Interoperability and initiatives to promote it are important for a number of interrelated reasons:

Safety and accountability benefits

- **Regulatory benefits:** Regulatory consistency can reduce forum shopping, and a regulatory "race to the bottom." Consistent documentation practices can make regulators' tasks easier (e.g., assessing a model's compliance with local laws and policies).
- **Safety and research benefits:** Consistent documentation practices can make it easier to compare, study, and evaluate models, which can support accountability across borders.
- **Benefits for the wider community:** Interoperability increases accountability across borders. Models operate and AI harms can manifest across borders. Greater accountability can increase trust and public confidence in AI by reassuring people that their rights will be respected and protected.
- **Promoting best practices and inclusion:** Interoperability initiatives provide an opportunity to promote best practices and inclusive policy development. If they are well-structured, they can set a baseline for policy, ensuring that international policy settings converge around best practices. This also supports the adoption of best practices in countries with less capacity to drive AI policy and can provide a vehicle to strengthen sustainable partnerships on AI. Including diverse voices in policy discussions in all jurisdictions is essential to realize this opportunity. Internationally, participation by Global Majority countries, in particular, should be supported in global forums, which

[J] The IPCC is the Intergovernmental Panel on Climate Change, the United Nations body for assessing the science related to climate change.

will set high-level objectives and baselines for interoperable policies.

- **Benefits in efficiencies and reduced environmental impacts:** Large models rely on significant compute capacity, both for training and inference. Policy and regulatory settings that permit access to models hosted across borders reduce the need to duplicate this infrastructure, reducing energy and water demands as well as reducing costs.

Innovation and economic benefits

- **Benefits for smaller AI actors:** Efficiency benefits will particularly benefit small and medium-sized enterprises ("SMEs") and startups, which are less able to absorb higher compliance burdens. It is also important for downstream developers in nations that do not host leading foundation model providers. Interoperability can mean that more internationally-sourced foundation models are available to downstream actors. Promoting consistent documentation also helps downstream actors to assess and thus choose between foundation models from other countries. That, in turn, means that local application developers have a greater opportunity to develop and use their expertise to build on these leading models to create niche products or products adapted to local needs.
- **Benefits for model providers and trade across borders:** Interoperable frameworks reduce compliance barriers, including costs, for foundation model providers. This makes it more efficient for new models to be released in multiple countries (and promotes faster access to new models).
- **Benefits for downstream actors in the AI value chain:** Consistent documentation can result in simplified compliance for downstream application developers, who integrate foundation models into AI systems. It also supports comparison of models, making it easier for developers to select the most appropriate models for their needs. It promotes access to models developed in other countries, supporting a global supply chain and allowing downstream AI systems and application developers in more countries to compete. Many foundation models are provided on a software-as-a-service model, so accessing them often requires access across borders.

Together, these benefits **promote innovation, support best practices, and promote access to the benefits of AI in more locations**. It should be noted that the safety/accountability benefits and the innovation/economic benefits are interrelated—for example, increased trust in AI systems will increase adoption; and research benefits will support technological development.

It is important to note that economic benefits from AI can be unevenly shared. AI policy, including on best practices for documentation, should be tailored to promote the broadness, inclusivity, and depth of these benefits.

The importance of interoperable AI policies is recognized both by national and multilateral policy initiatives (see Table 3 below).

Why should we care about documentation requirements being interoperable?

As discussed in the next section, documentation is a key element of policy and regulatory frameworks for foundation. It promotes accountability, safety, regulatory oversight, research, trade, and innovation. Documentation requirements are, therefore, a key focus area for interoperability efforts.

Documentation for foundation models

Documentation plays a key role in managing risk for foundation models. Therefore, transparency and documentation requirements are a common feature of foundation model policy frameworks. For the same reason, PAI's Guidance for Safe Foundation Model Deployment and Data Enrichment Transparency Template (recently released for public comment) contain recommendations for documentation practices and specific documentation artifacts across the AI lifecycle.

Foundation models have a number of features that make them more challenging to regulate than some other digital technologies:

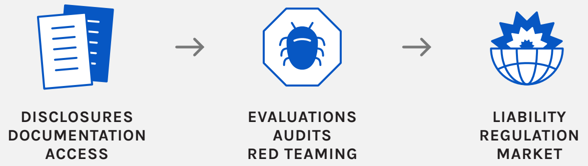
- **They can be used for many applications in many sectors.** This makes it difficult to regulate them by taking a traditional product safety approach, which addresses risk in particular sectors/use cases.
- **The most capable models are black-box systems.** It is not always possible to fully understand how they produce their results, and they can exhibit unexpected capabilities. Work is being done to address this, including by industry, but more work is needed.[K]
- **Large foundation models are trained on very large amounts of data, and finding a useful way to describe that data is difficult.** This makes it difficult to create detailed mandatory requirements for reporting on datasets.
- **The pace of innovation has been rapid, and the risks of frontier AI models and systems are not fully understood.** This increases the pressure to take policy action while making it difficult to know the precise form that action should take. It also makes it particularly difficult to create future-proof guidance or regulation. There is a difficult balance to achieve between creating sufficiently detailed frameworks while providing sufficient flexibility for those frameworks to evolve with the technical state of the art.
- **Foundation models are frequently fine-tuned and incorporated into AI systems by downstream actors in the AI value chain.** These downstream developers and deployers are required to ensure that the AI systems they develop are safe and legally compliant, which is challenging when they are building on complex and opaque foundation models.

Because of these factors, **documentation is a critical tool** for AI safety initiatives. It facilitates evaluations of models for potentially dangerous capabilities and provides a key input for downstream actors to identify and mitigate risks in the systems they develop.

Documentation is a key input for AI accountability. It supports audits and evaluations.[1] It supports downstream AI actors, including model adaptors, application developers, and model integrators to ensure they create safe AI models and systems. Documentation supports safety research.[2] It can also support civil society accountability and worker co-governance.

[K] Work is underway to decipher and understand the mathematical and scientific foundations of AI models (e.g., at DARPA, Anthropic, and ARIA) — but evaluating model capabilities remains a significant challenge.

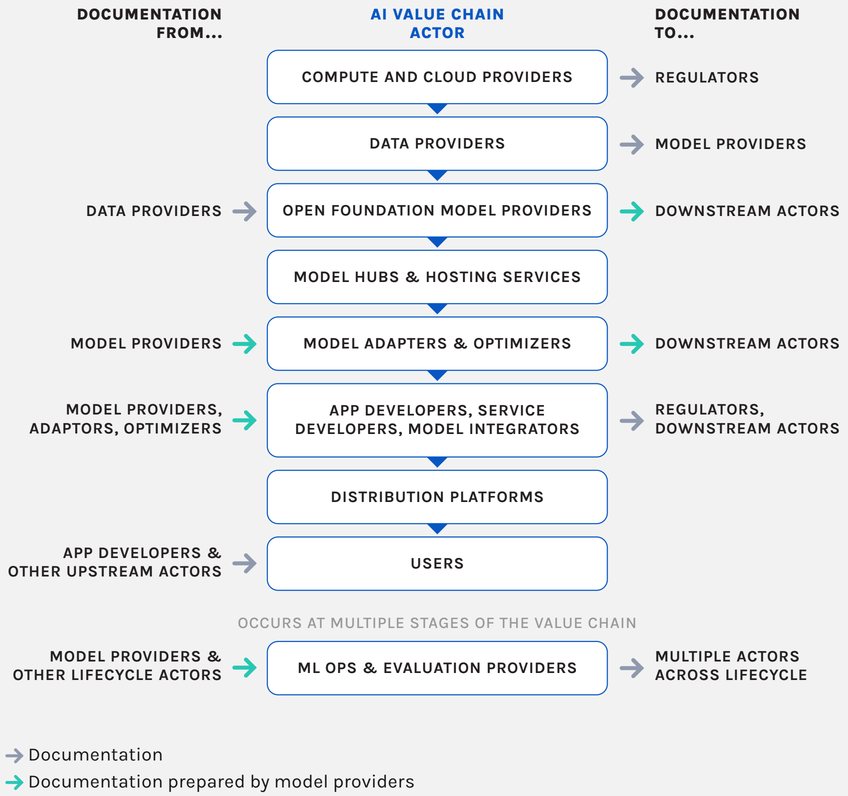
Documentation and the AI Accountability Chain



Source: NTIA AI Accountability Policy Report 2024

Transparency, including through documentation, is essential for national and international liability regimes to function effectively.[3] Robust and consistent documentation for foundation models helps allocate responsibility throughout the AI value chain. PAI's recent Risk Mitigation Strategies for the Open Foundation Model Value Chain explores the roles of different AI actors in mitigating risks. The providers and beneficiaries of documentation across the AI value chain are illustrated below.

Documentation and information flows through the AI value chain


DOCUMENTATION FROM...
AI VALUE CHAIN ACTOR
DOCUMENTATION TO...
COMPUTE AND CLOUD PROVIDERS
REGULATORS
DATA PROVIDERS
MODEL PROVIDERS
DATA PROVIDERS
OPEN FOUNDATION MODEL PROVIDERS
DOWNSTREAM ACTORS
MODEL HUBS & HOSTING SERVICES
MODEL PROVIDERS
MODEL ADAPTERS & OPTIMIZERS
DOWNSTREAM ACTORS
MODEL PROVIDERS, ADAPTORS, OPTIMIZERS
APP DEVELOPERS, SERVICE DEVELOPERS, MODEL INTEGRATORS
REGULATORS, DOWNSTREAM ACTORS
DISTRIBUTION PLATFORMS
APP DEVELOPERS & OTHER UPSTREAM ACTORS
USERS
OCCURS AT MULTIPLE STAGES OF THE VALUE CHAIN
MODEL PROVIDERS & OTHER LIFECYCLE ACTORS
ML OPS & EVALUATION PROVIDERS
MULTIPLE ACTORS ACROSS LIFECYCLE
→ Documentation
→ Documentation prepared by model providers

Adapted from PAI's Risk Mitigation Strategies for the Open Foundation Model Value Chain. It should be noted that this value chain was developed specifically for open foundation models. It has been adapted here to illustrate how documentation transfers information through the AI value chain.

Documentation should be a central focus of interoperability efforts

Policymakers impose documentation requirements/guidance for all the reasons discussed above. Documentation helps achieve desired policy outcomes (improving safety and accountability, supporting assurance and liability regimes, building trust, and promoting innovation). It also aids regulatory enforcement and helps inform further policy development.

While the importance of documentation is widely acknowledged, current practices vary widely, and **there is no consensus on best practices** for either the form or content of documentation artifacts. While a number of artifacts—including datasheets for datasets and model cards—are becoming quasi-standard, there is not yet a standard approach to completing these.[4] A number of difficult questions arise when considering what form documentation artifacts should take, including who the intended audience is (or should be). There may be risks to security, privacy or the release of trade secrets when sharing some documentation publicly. Policymakers, regulators, downstream developers and deployers, and independent researchers, are all likely to have different objectives when reviewing documentation and different levels of technical expertise to interpret it.[L]

The benefits of documentation are greatest when it is comparable across models and systems. This "facilitates understanding by familiarizing various stakeholders with a consistent process and format for documentation" and makes it "easier to judge relative performance, suitability, or impact."[5]

All these factors mean that interoperability for documentation requirements across policy and regulatory frameworks must be a core focus of efforts to harmonize foundation model guidance and regulation across borders.

While a greater degree of standardization in all documentation requirements would be useful, consultation participants indicated that **standardization of documentation for data used in training and validating foundation models** would be particularly helpful. Good data documentation helps downstream developers and deployers test and evaluate their products. It is an area where agreed-upon best practices remain elusive. It is also an area that presents some particular challenges, including the sheer size of datasets used in training the largest models; uncertainty about the interaction of privacy, copyright, and other laws with foundation model documentation; and the fact that foundation models may be deployed in a wide range of contexts; and the documentation needed to address risk in different contexts may vary.

Industry commitment to work, alongside wider stakeholders including from civil society, to improve documentation practices will be key in ensuring that common documentation practices are adopted by model providers. It was observed in our consultations that documentation practices currently vary widely, and best practices are not universally implemented. Internationally agreed requirements (in the form of interoperable policy frameworks) will assist in holding model providers accountable if the agreed practices are not followed.

[L] See, e.g., The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers

Foundation model documentation requirements

US, EU, UK, and multilateral policy frameworks

To explore how documentation guidance is being incorporated in current policy frameworks, we have compared leading policy frameworks from the US, EU, UK, and a number of multilateral bodies and initiatives. The reviewed frameworks are set out in Table 1 below. This table also summarizes what kind of provisions each framework contains about documentation for foundation models[M]—high-level transparency guidance, more detailed recommendations or requirements for documentation practices, and/or requirements for specific documentation artifacts. It also notes whether there are currently processes for building out each framework.[N]

Table 1: Frameworks reviewed in this paper

	A	B	C	D	E
Multilateral					
OECD AI Principles	●				
Seoul Frontier AI Safety Commitments	●	●			●
Hiroshima AI Process Code of Conduct		●	●	●	●
Council of Europe AI Convention		●			
Regional					
EU AI Act			●	●	●
National					
US AI Executive Order 14110			●	●	●
NIST AI RMF (with Gen-AI Profile)		●		●	●
UK AI White Paper and followup	●			●	

A: Contains high-level commitments to transparency

B: Requires/recommends documentation practices

C: Requires documentation artifacts

D: Further/more detailed provisions proposed or in development

E: Specifically addresses foundation models

We recognize that this is far from a complete mapping of all global policy initiatives addressing foundation models. Our aim for this stage of the work is to select a sample of frameworks from leading AI-developing jurisdictions, including:

- Examples of national, regional, multilateral, and international frameworks
- A mix of principles-based and binding frameworks
- Frameworks being developed to foster interoperability or a shared approach to responsible AI development

We chose a comparatively small number of frameworks to allow a more granular comparison and analysis of how they overlap and in what ways they may diverge as they are implemented and further developed.

It is important to recognize that most of the frameworks discussed in this report were developed by a relatively small number of countries (and multilateral bodies they are members

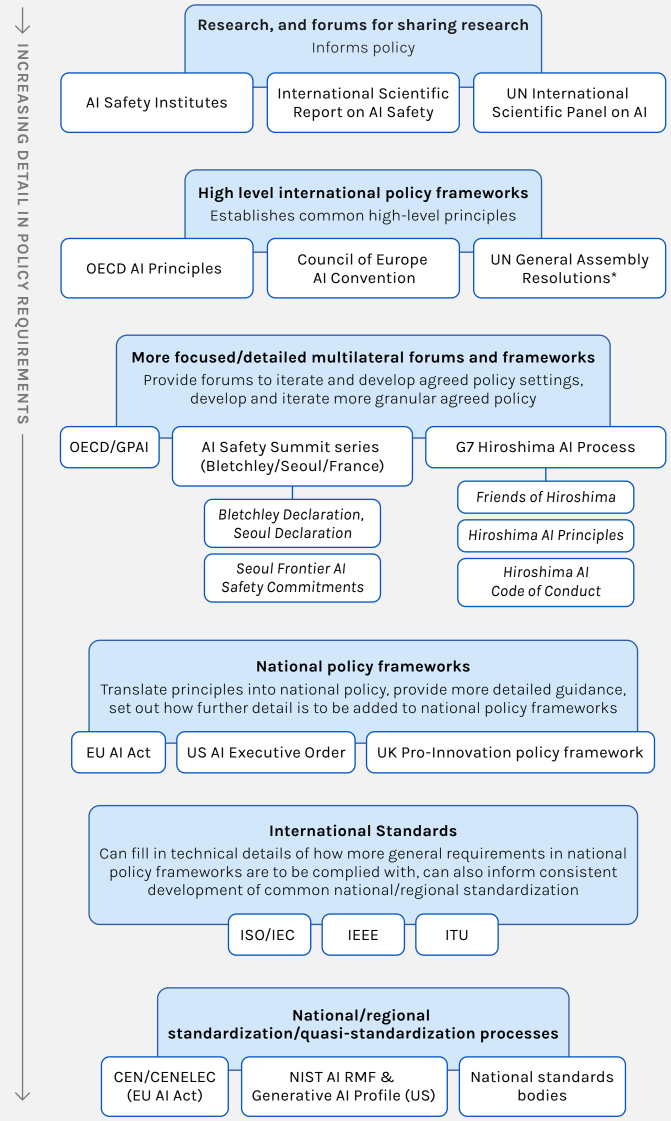
[M] Where frameworks do not contain specific provisions for foundation models, the table summarizes the requirements for AI systems more generally.

[N] This paper is focused on documentation for foundation models. Some of the earlier and high-level frameworks included below refer to AI systems, which will require differences in detailed documentation requirements, though the principles of transparency and accountability apply to systems and models.

of, like the G7). These countries are primarily in the Global North. These frameworks then are likely to embody the values and priorities of those countries. In selecting these frameworks as a starting point for examining interoperability challenges and opportunities, we acknowledge that efforts to achieve global interoperability will equally need to reflect the perspectives and priorities of the Global South.

It is to be expected that these frameworks would have different levels of detail about documentation requirements. **In an ideal scenario, these different policy tiers would interact to foster detailed interoperable policy frameworks at the national level, with internationally agreed high-level principles being developed through different multilateral fora for adoption in national frameworks.** National-level regulatory frameworks generally incorporate mechanisms for granular guidance/requirements to be developed in a way that is more easily amended than domestic legislation and adapted to highly technical and constantly evolving subject matter. This can be done through the adoption of subordinate or delegated rule-making; it can also proceed through reliance on or adoption of international standards. Ideally, the reviewed frameworks fit into this policy hierarchy outlined at right. This diagram illustrates how research and information sharing, as well as high-level agreements on principles or shared objectives, can inform the development of more detailed policy frameworks, which in turn can be further developed at the national or regional level. The national and regional frameworks, such as the EU AI Act, can contain significantly more detail than principles-based frameworks but still generally require further detail to be provided through mechanisms such as standards, codes of practice, or delegated rule-making.

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

* Interoperability and international cooperation are emphasized in two recent UN General Assembly resolutions on AI: Seizing the opportunities of safe, secure, and trustworthy artificial intelligence systems for sustainable development (document A/78/L.49) (March 11, 2024); Enhancing international cooperation on capacity-building of artificial intelligence (document A/78/L.86) (June 25, 2024). These resolutions reflect an important international commitment to promoting the beneficial use of AI, including through interoperability efforts, though they have not been included in the detailed comparison of frameworks in this document.

What if this chain is broken?

One key question that emerges from this view of the "chain" of policy-making, and the review of the policy landscape described in this report, is how interoperability is to be achieved if any of the policy-making layers are absent, or unable to fulfill their ideal role? In particular, the question of whether international standards will be able to play a key role in aligning foundation model policies is returned to in a later section of this report.

Documentation requirements in the covered frameworks

A general description of the frameworks reviewed in this report is given below, followed by a more detailed comparison of the documentation requirements or guidance they contain. As noted elsewhere in this report, these frameworks are drawn from a limited number of jurisdictions. While providing a starting point for considering how interoperability can be promoted, efforts to achieve globally interoperable frameworks will need to ensure global perspectives are fully incorporated.

High-level multilateral frameworks and initiatives

The **OECD AI Principles**, set out in the OECD Recommendation of the Council on Artificial Intelligence, state that AI actors should provide "meaningful information" to "foster understanding of AI systems" to "provide plain and easy-to-understand information on the sources of data/input, factors, processes and/or logic that led to" an output; to provide information to allow people to "challenge [an] output"; to "ensure traceability, including in relation to datasets, processes and decisions" over the AI lifecycle, and to "apply a systematic risk management approach to each phase of the AI lifecycle." The OECD has a number of other initiatives relating to AI Safety underway. It is currently developing sector-specific AI Due Diligence Guidance for Responsible Business Conduct.

The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (the "**COE AI Convention**") is an international treaty open to Council member states and other countries supporting the protection of human rights, democracy, and the rule of law. It will impose obligations on States Parties to take measures to ensure activities across the AI lifecycle are consistent with human rights, democracy, and the rule of law. It requires "appropriate" documentation to be made to achieve this end. **One interesting aspect of the Convention is its requirement that States Parties report regularly on their implementation of it**, which could, in time, provide a useful window into global regulatory approaches and developments.

More granular multilateral frameworks

Several multilateral initiatives have begun releasing more granular frameworks for advanced models.

The Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (the "**Hiroshima Code of Conduct**") was developed under the G7's Hiroshima AI Process. The process has been extended beyond the G7 membership through the Hiroshima AI Process Friends Group, currently comprising 52 countries and the EU. The Code is voluntary for providers of "the most advanced AI systems, including foundation models."

It includes requirements to document "measures to identify, evaluate and mitigate risks", to provide "regularly updated technical documentation", to maintain "documentation of incidents", and to publish "Transparency reports" with "meaningful information" that enables deployers/users "to interpret the model/system's output and to enable users to use it appropriately."

While the Code does refer to several documentation artifacts, **it does not provide any detailed guidance about the form or content** for these—for instance, it provides no guidance about what would constitute "meaningful information." The Code is intended to be a living document that will be "reviewed and updated as necessary." The OECD has released a pilot monitoring mechanism to assess voluntary compliance with the Code.

The Seoul Frontier AI Safety Commitments (the "**Seoul Commitments**") are another set of voluntary commitments launched at the AI Seoul Summit and endorsed by 16 model providers. They contain several transparency undertakings but do not provide any detailed guidance about the form or content for these.

National/regional level initiatives

USA

The **AI Executive Order** contains several reporting requirements for "dual-use" foundation model providers. It does not itself impose broader documentation requirements. The Order requires various federal agencies, including NIST, to undertake work focused on AI safety. It requires a large number of other initiatives to be taken by US federal agencies, including the preparation of the Generative AI Profile for the NIST AI RMF mentioned below.

The **NIST AI RMF** is similar to a process/management standard. It requires documentation of internal policies, various governance measures adopted, and mapping of risks. It recommends that many aspects of AI systems be documented, including a recommendation that "test sets, metrics, and details about the tools used during [AI Test, Evaluation, Validation and Verification ("TEVV")] are documented," as well as recommendations to document evaluations of the security and resilience of AI systems, transparency and accountability risks, privacy risks, fairness and bias, environment impacts, the effectiveness of TEVV metrics and processes adopted, measurement approaches for AI risks associated with deployment, and measurement results of AI system trustworthiness.

However, it does not recommend what tests and metrics should be adopted/documented, what documentation artifacts should be produced, and to whom they should be provided. It has recently been supplemented by the release of a Generative AI Profile for the AI RMF. The Profile contains guidance for the application of the NIST AI RMF to Generative AI, which it states "generally refers to generative foundation models."[6]

Together, the NIST AI RMF and Generative AI Profile contain significant guidance about documentation practices but do not contain detailed recommendations for specific documentation artifacts or detailed guidance about the form or content of artifacts.

European Union

The **EU Artificial Intelligence Act** sets out a comprehensive regulatory regime for safe and responsible AI in the European Union. Its most detailed provisions regulate "high-risk AI systems," which are AI systems intended to be deployed in particular contexts. However, it also contains specific provisions governing "General Purpose AI models" ("GPAI models")—these are defined[7] in a way that essentially overlaps with the concept of foundation models.

The EU AI Act provides more specific guidance than the other reviewed frameworks about particular documentation artifacts that should be produced for foundation models and has lists of what should be included in them. The Act provides for the development of harmonized standards as a means to demonstrate compliance for GPAI models (including documentation requirements). As an earlier step, it provides for the development of Codes of Practice addressing the same issues. In addition, the EU Commission has the power to adopt delegated acts about the thresholds for models posing systematic risks and to clarify the documentation requirements for GPAI models. The AI Office has the power to issue templates, including training data documentation, together with a range of oversight and enforcement powers.

United Kingdom

The UK is taking a sectoral approach to AI regulation. Its principal horizontal measure for AI safety is the creation of the UK AISI. Its existing approach to AI regulation is set out in a policy paper and a response to public feedback on that paper. The UK has not currently specifically regulated foundation models but has noted that it may be necessary to do so. The incoming Labour government has indicated that it intends to introduce legislation governing the most capable models and put the AISI on a statutory footing. The UK has endorsed five values-based principles based on the OECD's AI Principles.

Mapping documentation requirements across the frameworks

Documentation guidance/requirements under the reviewed frameworks are summarized in

Tables 2A , 2B and 2C below. Key findings include:

- **Documentation is a common feature of the frameworks**, though this is couched in various terms. Several high-level frameworks recommend providing certain kinds of information to various actors; some require recording and/or reporting of information, and some require the preparation of specific documentation artifacts.
- **The most commonly referenced artifacts** are (i) technical documentation, (ii) instructions for use, (iii) information about datasets, and (iv) incident reports.
- **However, there is little detail in most of the frameworks** about what should be included in each of these documents, and there is no guidance about the form each document should take.
- **This analysis suggests that there is an opportunity to develop standardized requirements** for some of the key documentation artifacts required across frameworks—provided that agreement can be reached about what the content of these artifacts should be.

Tables 2A, 2B and 2C below contain a comparison of documentation requirements across the in-scope frameworks. Specific documentation artifacts are shown in **red**. The principal documentation guidelines from PAI's Model Deployment Guidance are included as a comparator.

- **Table 2A** summarizes the kinds of documentation requirements in each of the reviewed frameworks, including whether they contain references to specific documentation artifacts or include guidance about when in the AI lifecycle documentation should occur.
- **Table 2B** compares those frameworks with more detailed documentation requirements across the AI lifecycle.
- **Table 2C** compares the less detailed frameworks which contain high-level statements of principle and/or do not address when documentation should be generated.

Table 2A. Comparison of documentation requirements across in-scope frameworks

	FRAMEWORK								
STAGE IN AI LIFECYCLE	PAI Model Deployment Guidance	EU AI Act	AI Executive Order	NIST RMF and Generative AI Companion	Hiroshima Code of Conduct	Seoul Frontier AI Commitments	COE Convention	OECD AI Principles	UK AI White Paper, AI Principles, Response
R&D	●	●	●						
Pre-deployment/ on deployment	● 🔴	● 🔴	●	●	● 🔴				
Post-deployment	●	● 🔴		●	● 🔴				
Across lifecycle	●			●	● 🔴				
Unspecified						🔵	🔵	🔵	🔵

● Documentation requirements for specific stage in the AI lifecycle

🔴 Specific documentation artifacts

🔵 General documentation requirements

Table 2B. Comparison of documentation requirements across in-scope frameworks

Specific documentation artifacts are shown in **red**. The principal documentation guidelines from PAI's Model Deployment Guidance are included as a comparator.

	FRAMEWORK				
STAGE IN AI LIFECYCLE	**PAI Model Deployment Guidance**	**EU AI Act**	**AI Executive Order**	**NIST RMF and Generative AI Companion**	**Hiroshima Code of Conduct**
R&D	**Pre-system card:** Planned testing, evaluation, and risk management procedures for foundation/frontier models prior to development. Including: • Intended training data approach • Responsible AI practices • Development Team • Written "safety case"	Notify EU Commission of models with systemic risk	Report dual-use models to Department of Commerce; report cybersecurity protections	N/A	N/A
Pre-deployment/ on deployment	Publicly report model impacts **"Key ingredient list":** including details of evaluations, limitations, risks, compute, parameters, architecture, training data approach, model documentation Disclose performance benchmarks, intended use, risks and mitigations, testing and evaluation methodologies, environmental and labor impacts **Downstream use documentation:** including appropriate uses, limitations, mitigations, safe development practices Share red-teaming findings	**Technical documentation:** including information about training, testing, and evaluations **Documentation for downstream developers:** including information about capabilities, limitations, and to aid downstream compliance **Public summary of training data**	Report red-teaming results to Department of Commerce	Multiple guidelines for documentation, including of: • Risks and potential impacts • Knowledge limits • TEVV considerations & tools • Measures of trustworthiness • Residual risks after mitigations • Model details • Data curation policies • Environmental impacts	**Technical documentation** **Transparency reports:** with "meaningful information" **Instructions for use** **Technical documentation** Documentation to include details of evaluations, capabilities/ limitations re: domains of use; risks to safety and society; red-teaming results
Post-deployment	Incident reporting Transparency reporting (frontier model usage and policy violations)	**Serious incident reports**	N/A	Incident and performance reporting Transparency reports with steps taken to update generative AI systems	Maintain "appropriate documentation" of reported incidents
Across lifecycle	Iteration of model development Provide documentation to government as required Environmental impacts Severe labor market risks Human rights impact assessments	N/A	N/A	Multiple guidelines to document processes and management systems	"Work towards" information sharing and incident reporting, including on: • Evaluation reports • Safety & security risks • "Ensuring appropriate and relevant documentation and transparency across the AI lifecycle" Document datasets, processes and decisions during development Regularly update technical documentation

Table 2C. Comparison of more general documentation and transparency requirements, at unspecified stages of the AI lifecycle

FRAMEWORK			
Seoul Frontier AI Commitments	**COE Convention**	**OECD AI Principles**	**UK AI White Paper, AI Principles, Response**
Publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use Provide public transparency on implementation of commitments, including on: • Risk assessments, effectiveness of mitigations, evaluation results • Risk thresholds • Approach to mitigations • Processes to follow if risk thresholds are met/exceeded	Countries ratifying the convention must have frameworks (such as national laws) that: • Contain documentation requirements that will allow people to seek remedies for human rights violations • Require developers to adopt measures to identify, prevent, and mitigate risk. These measures are to include documentation of risks and mitigations	Principles include: Transparency and Explainability: • "Provide meaningful information" to "foster understanding of AI Systems" • "Provide plain and easy-to-understand information on the sources of data/input, factors, processes and/or logic" • "Provide information [to] enable those adversely affected by an AI system to challenge its output." Accountability: • "Ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle"	Provide transparency and accountability, including "documentation on key decisions throughout the AI system life cycle"

Other features of the frameworks relevant to interoperability

In reviewing the in-scope frameworks, a number of additional factors emerge relevant to considering their current and potential future interoperability. These factors include:

- **Whether the frameworks are binding or non-binding, and whether they have any form of oversight or enforcement mechanism.** The most extreme example of inconsistency between frameworks would occur where binding frameworks contain divergent requirements. However, it is important to strive for interoperability between both binding and leading voluntary frameworks.
- **The coverage of the frameworks—that is, what kinds of models they apply to.** As discussed further below, the interoperability of frameworks is reduced if the coverage of those frameworks differs (or is unclear).
- **What institutions have functions to develop or oversee the frameworks.** Institutional cooperation is one mechanism that can promote interoperability, and mutual recognition of institutional functions, assessments, and other activities can also promote interoperability.
- **What mechanisms are in place to build out the frameworks.** E.g., through developing subordinate regulations or guidance and whether they support international standardization processes. These mechanisms give insight into where efforts to promote interoperability should be targeted.
- **Whether the frameworks prioritize collaboration and interoperability.**

These aspects of the frameworks are revisited in later sections of this report.

A comparison of these aspects of frameworks is given in Table 3 below. This shows that there are differences in the coverage of the frameworks and illustrates that the frameworks all envisage further work to provide more detailed guidance/requirements and are broadly aligned on the need for interoperability and the role of international standards in furthering that objective. These matters are discussed later in this report.

Table 3. In-scope frameworks: normative status, coverage/thresholds, reference to international standardization processes and collaboration/interoperability

Framework	Binding or Voluntary?	Coverage	Initial threshold	Institutions/ Oversight	Next steps	Commitment to cooperation/ collaboration	Commitment to standards
PAI Model Deployment Guidance	Voluntary	Foundation models (with guidance tailored according to three capability bands and four release strategies). The most stringent guidance applies to "paradigm-shifting or frontier" models	NA	N/A		Collaborate with cross-sector AI stakeholders re risk identification, methodologies, best practices, standardization	Development and adoption of standards
EU AI Act	Binding	General-purpose AI models (baseline requirements) General-purpose AI models "with systemic risk"	None (baseline requirements) 10^25 FLOPs (models "with systemic risk")	AI Office	Codes of Practice for GPAI due August 2025 Templates for training data (AI Office) Harmonized standards Delegated acts — thresholds for GPAI with systemic risk; documentation requirements	Mandates creation of AI Board, Advisory Forum; multistakeholder participation in development of Codes of Practice and harmonized standards	EU harmonized Standards—though EU committed to adopting international standards where possible[8]
AI Executive Order	Partly binding	"Dual-use foundation models"	10^26 FLOPs (10^23 FLOPs for models trained on biological sequence data)	Dept of Commerce (for reporting requirements)	Various, including: OMB materials for federal procurement Copyright guidance Dept of Commerce can change threshold for dual-use model reporting	Under EO, NIST released plan for global engagement on AI standards; Secretary of State is developing Global Development Playbook; EO contained several consultation requirements	NIST is required to develop standards Under EO, NIST has released plan for global engagement on promoting and developing AI standards
NIST RMF and Generative AI Companion	Voluntary *While the NIST AI RMF and Generative AI Profile are not binding, they will be picked up through US federal procurement guidance.*	AI systems (NIST AI RMF) Generative foundation models (Gen-AI Profile) *The Profile applies to "Generative AI," and notes that "for purposes of this document, GAI generally refers to generative foundation models."*	N/A	N/A	NIST/the NIST AISI have a broad work plan including developing tools, evaluations, metrics	Several references to collaboration e.g. with external researchers, industry experts, and community representatives about best risk measurement and management practices NIST is committed to collaboration/cooperation, e.g. through AISI Consortium and pending Network of AISIs	Contains references to considering relevance of standards (including NIST frameworks) NIST will continue to align AI RMF with international standards[9]
Hiroshima Code of Conduct	Voluntary	"The most advanced AI systems, including the most advanced foundation models and generative AI systems"	N/A	OECD (monitoring mechanism under development)	COC to be iterated by G7 HAIP OECD developing monitoring mechanism	Across sectors, including on research to assess/adopt risk mitigations, document incidents, and share information with the public to promote safety	Advance development and adoption of standards
UK AI White Paper, Consultation Response	Voluntary (guidance for sectoral regulators)	AI systems; generally a sectoral approach Initial focus of UK AISI on advanced systems Planned laws for "the most powerful AI systems"[10]	N/A	AISI	Intention to legislate announced re advanced models, and to place AISI on statutory footing	Focus on collaboration across government, stakeholder groups, and internationally	Support for work on assurance techniques and technical standards
OECD	Voluntary	AI Systems	N/A	N/A	OECD developing Due Diligence Guidance (DDG) for AI under OECD Responsible Business Conduct (RBC) guidelines	OECD convenes the Network of Experts	Governments should promote standards development
Seoul Frontier AI Commitments	Voluntary	"Frontier AI" — "highly capable general-purpose AI models or systems that can perform a wide variety of tasks and match or exceed the capabilities present in the most advanced models"	N/A	N/A	AI Action Summit February 2025 (France) AI Safety Science Report to be published at AI Action Summit	Information sharing, collaboration on safety research (Seoul AI Principles)	Contribute to/ take account of international standards

Interoperability of documentation requirements across the frameworks

What we see and what we don't see

From the analysis of the frameworks in the preceding section, it is apparent that there are no current significant interoperability challenges, though not all frameworks call for the same artifacts. There is also a lack of clarity (for instance, in the Hiroshima Code of Conduct) about which information is to be included in which artifact, potentially leading to a diversity of practices. The documentation recommendations/requirements are, therefore, not in conflict just yet. This is perhaps unsurprising, given many of the frameworks, particularly the international frameworks, do not contain significant levels of detail. However, in the months ahead, as further detail is developed and driven forward (e.g., through the forthcoming development of EU Codes of Practice, documentation work in the US under the NIST AI Safety Consortium, and potential further development of the G7 Code of Conduct), ensuring that issues do not emerge will depend on efforts made to coordinate these initiatives.

Interoperability efforts should not only be made to avoid policy divergence/fragmentation, but also to ensure that a consistent baseline of good practice, building on existing and future research about documentation practices, is adopted and built into the various frameworks to drive best practice across borders.

Interoperability efforts should not only be made to avoid policy divergence/fragmentation, but also to ensure that a consistent baseline of good practice is adopted.

What documentation requirements do we see right now?

While the frameworks vary in the degree of specificity with which they address documentation, there is convergence in the matters they recommend be documented. These include: information about models such as algorithms and architecture; training data; model capabilities; testing and evaluations that have been conducted and the outcomes of these; and appropriate uses. However, none of the frameworks yet provide significant detail about how to go about implementing this guidance.

In some cases, the frameworks may align on what should be documented but not on whether that should be included in a specific artifact. They also differ in whether certain documentation should be made publicly available. Documentation of training data is a significant example in both cases.

Risk and opportunity 1: Divergence in details

Several of the frameworks—the EU AI Act and the Hiroshima AI Code of Conduct—recommend/require the preparation of **specific documentation artifacts**. Both these frameworks refer to the preparation of **technical documentation** and **documentation for downstream developers**. There are differences between the frameworks, however:

- The EU AI Act requires the preparation of a **public summary of training data**, while the Code of Conduct requires datasets to be documented but does not require this to be done in a public-facing or standardized artifact.

- The Code of Conduct recommends the preparation of “**Transparency Reports**” but is vague about their content.
- The EU AI Act requires the documentation of **serious incidents** and reporting to the EU AI Office, as well as national bodies; the Code of Conduct recommends the recording of incidents but does not address reporting by model providers.

In general, the Code of Conduct contains less detail than the EU AI Act about what information should be recorded in each artifact. The comparison of these frameworks in Tables 2A, 2B and 2C suggests that at this stage it could be possible to develop these frameworks in a mutually consistent way at a high level. Given both the AI Act and the Code envisage that further work will be required to add specificity to the requirements they contain, collaboration between the G7 and EU to define more clearly “best practices” and translate them into concrete requirements under the Act and the Code will be required to ensure this work is done in a coherent way. As discussed further below, work on the EU Codes will commence shortly, so this is a matter of some urgency.

Risk and opportunity 2: Binding and voluntary frameworks should align

Most of the frameworks reviewed are not binding. Currently, only the EU AI Act and the US AI Executive Order have binding provisions affecting foundation model providers (and only the EU AI Act contains a detailed binding regime).[O] This is likely to continue at the international/multilateral level, though these frameworks may serve to inform the development of more detailed frameworks at the national level. At the national level, it is possible that further legally binding requirements will be introduced over time within the countries reviewed in this paper and beyond.[P] It is also likely, however, that some countries will continue to take a voluntary approach to at least some aspects of foundation model policy. As discussed earlier in this paper, interoperability promotes multiple values including transparency, accountability, and trust in AI systems, as well as wider access to leading models across borders, regardless of whether the frameworks are backed by legal sanction. **Interoperability efforts must include efforts to promote alignment between both binding and non-binding frameworks.**

Risk and opportunity 3: More detailed guidance is needed—presenting both an opportunity and a challenge

It is clear from the comparison of the frameworks that even in the more detailed frameworks, such as the EU AI Act and the NIST AI RMF, there is a **lack of guidance about the form and content** of documentation artifacts for foundation models.

As discussed above, **this guidance is needed**. In the case of the EU AI Act, it is under development in the form of Codes of Practice for GPAI models. Developing this more detailed guidance is where the biggest challenge to interoperability will lie moving forward—as without coordination, different jurisdictions may take different approaches. The more forums are tasked with developing more detailed requirements, the greater the risk of this divergence.

[O] The Council of Europe AI Convention will be binding on States Parties when it enters into effect, but will not itself impose obligations on model providers.

[P] The UK announcement of further proposed legislation is an example. Legally binding regimes have been proposed in a number of other jurisdictions including Canada and Brazil.

However, to provide useful guidance to model providers and, therefore, to ensure the benefits of foundation model documentation (including promoting accountability) are realized, more detailed provisions are required. The primary interoperability challenge will be ensuring that this more detailed guidance under various national frameworks is developed in a coherent way.

Risk and opportunity 4: Interoperability with other (non-AI-specific) legal requirements

Several consultation participants noted concerns that as more detailed documentation requirements are developed, **tensions, and potentially direct inconsistencies, could arise** with other regulatory regimes that are not specifically AI-focused. For example, it was argued that the requirement to document certain categories of personal data could conflict with data minimization requirements. Documentation of copyright material was another example given. There was no consensus around these issues, however. While consideration of these non-AI focused laws is outside the scope of this paper, these concerns indicate policymakers need to consider consistency with broader legal requirements when formulating horizontal frameworks for AI models and systems and to engage in broad multistakeholder dialogue when doing so.

Recommendations

Looking ahead, what steps should policymakers take to drive a baseline for good practice and accountability while improving interoperability?

Consultation participants identified a number of factors that can promote interoperability between policy frameworks.

Principles to adopt

- **Start early. It is important to think about interoperability at an early stage of policy development.** Otherwise, incompatibilities can become entrenched and difficult to resolve at a later time. Working towards interoperability will necessarily be an ongoing, iterative process, as technological developments require revisions and updates of policy settings.
- **Agreeing on common principles and high-level frameworks** promotes interoperability when more detailed policy frameworks are developed at the national or regional level.
- **Policy consensus is built on shared understanding.** As noted elsewhere in this report, there is no agreement on some key issues underpinning foundation model regulation. This includes how to assess model capabilities, risks, the effectiveness of mitigations, and how to determine acceptable risk thresholds. There is also, as yet, no agreement on what either baseline or best practices for foundation model documentation should

be. A key element of working to promote interoperability is collaborating on AI safety research, including information-sharing, to enable the development of consensus about best practices, which should be built into policy and regulation.

- **Ensure that efforts to promote interoperability do not lead to convergence on a "lowest common denominator" for policy.** Rather, these efforts should be used as an opportunity to align on good practices that will advance the interests of society, including by addressing all relevant risks. This principle is critical to ensure that we foster good practice, as interoperability established on harmful practices could grow risks within and across countries. Interoperability should be pursued when a good baseline of practice has first been achieved.
- **Incorporating multistakeholder perspectives in policy-making is essential.** This includes at the national and the international levels, as well as in forums working to promote international interoperability. Many of the policy forums discussed in this paper feature significant industry representation but lack adequate representation from civil society and the Global South.
- **Filling the gap while policy and regulation are developed.** Policy development, and in particular, regulation, takes time. There is a role for voluntary frameworks that establish best practices led by multistakeholder bodies in promoting responsible foundation model development while policy frameworks continue to be developed. Examples include PAI's Guidance for Safe Foundation Model Deployment.

While interoperability is important, it isn't necessarily a goal requirement in every area. For this paper, we focus on the need and opportunity for policymakers to build interoperability specifically for documentation, given the key role it plays in promoting accountability and responsible foundation model development and deployment—and, therefore, the importance of achieving a baseline of good documentation practices across borders.

RECOMMENDATION 1

Coverage and thresholds should be aligned to the extent possible

Review of the frameworks considered in this report indicates that several have broad coverage, applying to all AI systems. A number, however, make specific provisions for foundation models or some subcategories of them. In particular, they provide special provisions for highly capable models. The terminology, definitions, and **thresholds** for these models vary between the frameworks.[Q]

[Q] Thresholds "describe AI capabilities beyond which an AI system is deemed to pose too much risk." These can be based on the compute used to train models or other factors. While a number of current frameworks adopt compute-based thresholds, the appropriateness of this approach is disputed.

Why might this be a risk to the interoperability of documentation requirements? Because agreeing on a common definition for the subcategory of powerful foundation models warranting additional controls is a key first step to promoting interoperability between these frameworks.

Review of the frameworks suggests that the differences in thresholds are driven in part by questions of model capability and partly by different frameworks focussing on managing different risks. The various coverage and thresholds are summarized in Table 3 above.

The **COE AI Convention** and the **OECD AI Principles** apply generally to **AI systems**. They contain some general recommendations relevant to foundation models but do not contain specific provisions directed to them or specific guidance tailored to more powerful foundation models (or systems).

The **EU AI Act** includes two sets of requirements for GPAI models. The first applies to all GPAI models.[11] The second applies to **GPAI models**[R] **"with systemic risk."**[12] The principal initial definition for models with systemic risk is those models that have "high-impact capabilities."[13] Models have "high impact capabilities" if they have "capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models."[14] Models will be taken to meet this criterion if the computation used in their training is more than 10^{25} FLOPs. In addition, the EU Commission may designate a GPAI model as having systemic risk, taking into account a number of criteria specified in an Annex to the Act.[15] The EU Commission has the power to amend these criteria. In doing so, it will be guided by the overarching definitions in Article 3. The substance of these is that models will present "systemic risk" if they have capabilities matching or exceeding those of the most advanced GPAI models and, by virtue of those capabilities, present a risk of having "**a significant impact on the EU market due to their reach or impact on health, public security, fundamental rights, or society as a whole.**"[16]

[R] The definition of General Purpose AI models is substantially the same as the accepted definition of foundation models. The term "foundation model" was coined by the Stanford Institute for Human-Centered Artificial Intelligence's (HAI) Center for Research on Foundation Models, as "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks." Under the EU AI Act, a General-purpose AI model is an "AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks … and that can be integrated into a variety of downstream systems or applications" (Art. 3(63)).

The **AI Executive Order** contains a raft of provisions relating to the use of AI by the US federal government. In addition, it identifies "**dual-use foundation models**" as warranting specific governance measures. These provisions are not limited to government use. A dual-use foundation model is:

> *an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a* ***serious risk to security, national economic security, national public health or safety, or any combination of those matters****, such as by:*
>
> *(i). substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;*
>
> *(ii). enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or*
>
> *(iii). permitting the evasion of human control or oversight through means of deception or obfuscation.*[17]

The AI Executive Order contains an initial threshold for models deemed to meet this definition.[18] [S] Like the EU AI Act, it is based on the computation used to train the model. However, the Executive Order contains a higher threshold of 10^{26} FLOPs.[T] The Secretary of Commerce is authorized to amend this threshold.[19]

[S] This threshold applies to the (limited) reporting requirements for model providers in the AI Act.

[T] A lower threshold is set for foundation models primarily trained on biological sequence data: 10^{23} FLOPs.

As set out above, the initial thresholds for regulating advanced foundation models under the

EU AI Act and the AI Executive Order are different (10^{25} vs 10^{26} FLOPs). More than this, however, the underlying definitions of "dual-use" foundation models and GPAI models "with systemic risk" in the EU AI Act and the AI Executive Order are not the same and are tied to specific types of risk posed by models. While there is significant overlap between these definitions, they are not identical. This means that when the compute-based initial thresholds in the EU AI Act and the AI Executive Order are revised, they may diverge further; and the measures required to address the risks posed by the models may, therefore, also diverge.[U]

Other frameworks also apply to a subcategory of powerful foundation models. They use a variety of terminology to refer to these:

- The Hiroshima Code of Conduct applies to "**the most advanced AI systems**, including the most advanced foundation models and generative AI systems."
- The Seoul Frontier AI Safety Commitments apply to "**Frontier AI**", defined to be "highly capable general-purpose AI models or systems that can perform a wide variety of tasks and match or exceed the capabilities present in the most advanced models."
- The UK has announced it intends to introduce legislation governing "**the most powerful AI systems**"; it is expected this will include the most advanced foundation models.

These definitions for powerful foundation models share significant overlap. Most are tied to models that match or exceed the "state of the art" (though do not specify whether this refers to the state of the art at the time the frameworks were introduced or if the scope of the models captured is intended to evolve with the state of the art).

Agreeing on a common definition for the subcategory of powerful foundation models warranting additional controls is a key first step to promoting interoperability between these frameworks. First, it will provide certainty to model providers, civil society, academia, and the public about what requirements industry needs to meet for a proposed model (i.e., whether a model is covered by a particular framework). More importantly, the measures required to manage risks posed by powerful models are likely to depend on the capabilities of the in-scope models and the categories of risk that the frameworks are intended to address. That is, agreeing on a common definition and thresholds for the models covered by policy frameworks may flow through to greater alignment between the frameworks, including in relation to documentation requirements. Further, one key matter that is commonly required in documentation is information about the testing and evaluations that models are subjected to; common thresholds will, therefore, again feed into more comparable documentation.

The AI Summit series and the Hiroshima AI Process are both forums that could usefully advance international collaboration on this issue. The announced Network of AI Safety Institutes (discussed further below) would also be well-positioned to contribute to this task. Finally, the OECD has recently announced an initiative investigating this issue. The OECD's broad membership, its recent integration with the Global Partnership on AI, and its current collaboration with the Hiroshima Process mean it is particularly well-placed to influence policy consensus on this topic.

[U] Revision of these thresholds will be required as techniques to assess the capabilities and attendant risks of advanced foundation models improve. In particular, there are criticisms of the current reliance on compute as a proxy for risk (see, e.g., On the Limitations of Compute Thresholds as a Governance Strategy, though others have suggested that despite being imperfect, there may be a place for compute (potentially together with other metrics) in setting thresholds — see, e.g., Frontier AI Regulation: Managing Emerging Risks to Public Safety

RECOMMENDATION 1

National governments and the EU should prioritize cooperation in identifying thresholds for identifying which foundation models require additional governance measures, including through engaging with the OECD's work on this issue. The AI Summit Series could also be used to take this forward.

It should be noted that the policy frameworks (or those parts of them) that incorporate the thresholds discussed above apply to a small number of "frontier" or "advanced" foundation models. Alignment between these frameworks is an important objective as they are currently an area of significant policy activity, and these advanced models present particular risks. As set out in PAI's Model Deployment Guidance, appropriate measures to identify and mitigate risks should be implemented for all foundation models, not only those that meet or exceed the state of the art.

RECOMMENDATION 2

A potential first step to interoperability? Aligning EU Codes of Practice and future iterations of the Hiroshima Code of Conduct

A key step in working towards interoperable policy frameworks is ensuring that interoperability is considered at an early stage and, therefore, by policy first-movers. Two of the more advanced frameworks at the national/regional and international/multilateral level are the EU AI Act and the Hiroshima Code of Conduct. Both of these envisage further development (in the case of the EU AI Act, these processes are underway). Both include provision for oversight (in the case of the EU, the EU AI Office has extensive oversight and enforcement powers. In the case of the Hiroshima Code of Conduct, the OECD has launched a pilot to monitor the application of the Code). Seeking alignment between processes to build out these frameworks, therefore, presents an opportunity to develop a foundation for interoperability as further national frameworks are developed.

The most developed policy framework of those reviewed is the EU AI Act. The product of several years of negotiation, it is now law (though not all of its provisions are in effect yet), and establishes concrete processes for creating detailed guidance for foundation model documentation. The most significant of these in the near term are the requirements for the development of Codes of Practice for GPAI model providers, which will need to cover the documentation requirements in the EU AI Act (including technical documentation, downstream use documentation, and serious incident documentation for GPAI models with systemic risks). While these will not be binding, providers will be able to "rely" on them to demonstrate compliance with the AI Act's requirements for GPAI models.[20]

The most developed international framework to date is the Hiroshima Code of Conduct. While it remains high-level, it recommends the creation of several documentation artifacts, and the G7 has announced that it will be reviewed and updated as necessary. It is the product of the Hiroshima AI Process. It has also affirmed that the G7 will continue to work to promote interoperability for AI governance. While the G7 has a narrow membership, the Hiroshima

AI Friends Group now has over 50 members. This makes the Code of Conduct a promising vehicle to promote interoperability. One promising suggestion is that **the G7 continue to build out the Code of Conduct to align with the Codes of Practice being developed under the EU AI Act**—or even for the Code of Conduct to be developed to the point that it could **constitute a Code of Practice under the EU AI Act**. That would be a significant step towards promoting international interoperability. Even if that goal is not practicable in the near term, developing the Hiroshima Code of Conduct with a goal of informing or aligning with the initial EU Codes of Practice would be a valuable step towards interoperability. To realize that potential in a manner that is based on best practice, globally relevant, and informed by individuals with sociotechnical expertise, efforts are needed to ensure meaningful participation in these processes by civil society organizations and geographically diverse stakeholders, including non-members of the EU and G7.

Aligning the EU AI Act and the Hiroshima Code of Conduct would, of course, be only a first step towards global interoperability. It would nevertheless be valuable for several reasons.

- As early movers in policy development in this space, the EU and the G7 are likely to influence subsequent policy initiatives. This is an opportunity while also being a risk if the EU and G7 do not take proactive steps to consult a wider set of countries within and outside of the Friends of Hiroshima grouping.
- The EU and the G7 include as members national governments hosting the majority of the world's leading foundation model developers and deployers. That makes policy development in those countries particularly important. It also means policymakers may have greater access to inputs from leading model providers to inform policy development.
- With the EU's large membership and the Friends of Hiroshima grouping, as well as the G7's ongoing collaboration with the OECD, policy alignment between these processes will lead to alignment between a substantial number of countries.

Together, these factors support promoting interoperability between documentation requirements in the EU AI Act and under the Hiroshima Code of Conduct in consultation with a wider set of countries. This will increase policy alignment between a large number of countries in which access to the inputs needed to create good policy is highest, and where that policy will have a significant impact in the near term.

RECOMMENDATION 2A

The G7 Presidency should continue developing the Hiroshima Code of Conduct into a more detailed framework, specifically to provide more detail about thresholds, relevant risks, and the form and content of documentation artifacts. This work should be a focus of Canada's G7 Presidency in 2025, including to align closely with the Codes of Practice development timeline. In doing this, it should seek input from foundation model providers, civil society, academia, and other stakeholder groups equally.

RECOMMENDATION 2B

In developing and approving the initial Codes of Practice under the EU AI Act, participants in the development process, the AI Office, the AI Board, and the EU Commission should adopt as a key objective, to the extent practicable, interoperability with other leading frameworks.

Recommendations for institutions playing different roles to support interoperability and achieving best practice

RECOMMENDATION 3

Standards

Traditionally, one of the major levers for the harmonization of international regulation for technology, as well as technical interoperability, have been standards developed by international SDOs. There was widespread support among consultation participants, as well as in the literature, for standards to play this role in advancing international interoperability for foundation model policy. However there are a number of potential barriers to overcome for that to occur.

In the field of AI, leading international standardization efforts include those led by ISO/IEC JTC 1/SC 42, the ITU, and the IEEE. Meanwhile, In the EU, the development of harmonized AI standards is being led by CEN-CENELEC JTC 21. A significant number of AI standards have been developed and are in development, though to date, there is little specific treatment of foundation models. There is potential for this to change and for these standards to become more prominent from 2025 onwards, building on the Code of Practice that will have been developed under the EU AI Office by then.

Consultation participants strongly endorsed the role that standards can play in promoting interoperability, including standardization of documentation requirements for foundation model providers.

Standards are not directly binding but can be given legal status under national laws.[21] They can also play a role as a reference point to establish a standard of reasonable conduct under general liability frameworks. International standards have traditionally been a key part of the assurance ecosystem—providing the benchmarks for certifications and audits that permit digital technologies to be authoritatively assessed to be fit for deployment across borders.

Consultation participants strongly endorsed the role that standards can play in promoting interoperability, including standardization of documentation requirements for foundation model providers.

There are several factors that could make it more challenging for international standardization processes to play this role in the context of foundation models. These include:

- **Standards rely on scientific knowledge/consensus.** The necessary degree of consensus does not yet exist for managing all risks associated with foundation models.[V]
- **Multistakeholder participation is essential for developing policies** to manage

[V] NIST's recent Plan for Global Engagement on AI Standards (2024) sets out priority areas for AI standardization, noting areas where there is insufficient scientific consensus for work to proceed (at pp. 9-14).

foundation model risks.[W] While ISO/IEC, IEEE, and CEN-CENELEC standardization processes all include multistakeholder engagement, in practice, it can be challenging for representatives from civil society, academia, and the Global South to meaningfully participate.[X] Input from a wide range of perspectives is particularly important for foundation model policy development. Barriers include:

  - SDOs are membership-based bodies, generally requiring membership fees; and both draft and final standards are behind paywalls.
  - Engagement in SDO processes is time-consuming and requires significant technical expertise. This is a particular challenge for some civil society organizations.

- **There are some issues on which standards bodies are currently less well-positioned to inform AI policy given their lack of diversity and focus on sociotechnical elements of AI.** This is most notably the case in questions involving impingements on fundamental rights, labor, environmental impacts, and supply chains more broadly. For example, setting acceptable thresholds for unfair bias and discrimination involves policy questions as well as sociotechnical expertise. It also raises questions of legitimacy, given that acceptability should probably be decided in democratic institutions. Standards bodies will need to address these issues to properly build out good practices for documentation (and it is possible some of these policy questions will need to be addressed by other bodies or processes to be fed into standardization efforts).
- **International SDO standardization processes take a long time.** This presents a dual challenge: given the pace of innovation, policy development is needed in the near term; and policy frameworks are already in place/being developed, so there may be difficulties with standardization processes keeping up. In general, this is a less significant issue for management/process standards, which do not need to be iterated as frequently. But it does mean new standards cannot easily fill a need for immediate guidance.
- **CEN-CENELEC endeavors to align EU standards with international standards through formalized technical cooperation processes.** However, it is likely that EU AI standards will diverge from international standards in some respects. Harmonized standards play a specific role under the EU AI Act. Compliance with these standards will bring a presumption of conformity—that is, deemed compliance with the AI Act. That means that harmonized standards must be tailored to the requirements of the AI Act, including its definition of risk and its treatment of fundamental rights. It is not yet clear whether these challenges would apply to potential standards addressing foundation model documentation requirements, and the development of harmonized standards for GPAI models is likely some way off. In the immediate term, the role of harmonized standards will be filled by the Code of Practice, although the Code of Practice might not have the same "formal" presumption of conformity nor the formal ties to the global standards structure (e.g., the work of the ISO). While the Code of Practice is intended to play a similar implementation role to harmonized standards under the EU AI Act in the immediate term, there may be equivalent challenges in developing the Code in a way that is consistent with any future international standards.

[W] This is recognized in the frameworks reviewed in this paper, as reflected in Table 3.

[X] This continues to be so despite efforts to address the issue.

Despite these challenges, standards remain an important tool for developing interoperable AI policy frameworks, and they should be developed to manage foundation model risks, even if they cannot address all aspects at this stage.

RECOMMENDATION 3A

To support the development of standardized documentation artifacts such as dataset documentation and technical documentation, Standards Development Organizations should ensure that their processes are informed by appropriate sociotechnical expertise and diverse perspectives, as well as required resources. To that end, SDOs, industry, governments, and other bodies should invest in capacity building for civil society and academic stakeholders to engage in standards-making processes, including to ensure participation from the Global South. That could include engaging in more active outreach and providing financial and logistical support. This is critical to ensure multistakeholder, sociotechnical and global expertise informs these processes. Governments should consider mirroring initiatives such as the UK's AI Standards Hub to achieve this goal.

RECOMMENDATION 3B

The development of standardized documentation artifacts for foundation models should be a priority in AI standardization efforts.

RECOMMENDATION 4

Ongoing collaboration, research, and the science of AI Safety

A recurring theme throughout the research and consultations for this report was that there is not yet consensus about what the best practices are for foundation model documentation. This lack of agreement makes it difficult to develop detailed policy frameworks at the national level, let alone internationally interoperable frameworks. This lack of clarity about best documentation practices derives in part from the need for further research on foundation model capabilities, risks, and mitigations; and the need for agreed tools, evaluations, and benchmarks for these.[Y]

This suggests that **collaboration on research could lead to shared understandings** of foundation model capabilities, risks, and mitigation measures; which could in turn provide a foundation for the **development of agreed best practices** for documentation to be incorporated into policy frameworks. This could be a significant driver for interoperable policy development.

The need for further research on the science of AI Safety, and international, multistakeholder collaboration in that endeavor, is a notable feature of the policy frameworks and initiatives outlined in Table 3 above. A number of international initiatives are currently underway to further this work.

[Y] Standardized documentation for models will include documentation of capabilities, risks and mitigations, and evaluation outcomes. Other challenges to agreeing on best practices for documentation include that foundation models can be deployed in many contexts, and the documentation needed to build, test and evaluate downstream AI systems varies with the context of use.

In May 2024, the Interim International Scientific Report on the Safety of Advanced AI was released at the AI Seoul Summit. The final report is anticipated before the AI Action Summit in France in 2025. There would be utility in continuing this reporting process. One option for this would be to continue the preparation of periodic reports under the aegis of the AI Summit process. Another option would be for this process to be transferred to a genuinely international initiative.[Z] The recent UN Global Digital Compact, endorsing the recommendation of the UN High-Level Advisory Body on AI, includes a commitment to establish an International Scientific Panel on AI, tasked with issuing issuing both annual and ad hoc reports surveying AI capabilities, opportunities, risks and uncertainties, identifying areas of scientific consensus and areas where more research is needed, and discussing emerging issues and gaps in the governance landscape. This body could play a valuable role in advancing consensus on the science of AI Safety to inform coherent policy development internationally. A particular benefit of the International Panel is that it will provide an avenue for a broader range of global perspectives to be taken into account. This is particularly important given the fact that many of the currently leading frameworks (including the ones discussed in this report) originate from countries in the Global North (or multilateral bodies/initiatives with memberships largely drawn from the Global North).

RECOMMENDATION 4

International collaboration and research initiatives should prioritize research that will support the development of standards for foundation model documentation, including dataset documentation and technical documentation. Documentation is a key feature of foundation model policy requirements, and common requirements for artifacts will directly improve interoperability. It will also make comparisons between models from different countries easier, promoting accountability and innovation.

[Z] While the writers, expert advisory panel, and science committee of the Interim International Scientific Report include geographically diverse representatives, the AI Summit process under which it is auspiced is not yet a fully international one.

RECOMMENDATION 5

The AI Safety Institutes and the Network of Institutes

The US, UK, and EU have given key roles in overseeing foundation models to new institutions. The UK and the US have established AI Safety Institutes, with functions centered on advancing the science of AI Safety, developing safety tools and techniques, and conducting evaluations of advanced models. In the EU, the AI Office has similar functions in addition to a wider suite of regulatory powers.

Table 4: AI Safety bodies in the US, UK and EU

	US NIST AISI[22]	UK AISI[23]	EU AI Office
Mission/Goals	• Advancing the science of AI safety • Articulating, demonstrating, and disseminating the practices of AI safety • Supporting institutions, communities, and coordination around AI safety.	"To minimise surprise to the UK and humanity from rapid and unexpected advances in AI."	Supporting "the development and use of trustworthy AI, while protecting against AI risks"[24]
Functions/ Activities	• Technical research on safety guidelines and technical safety tools and techniques • Conduct pre-deployment TEVV of advanced models, systems, and agents to assess potential and emerging risks • Conduct TEVV of advanced AI models, systems, and agents to develop scientific understanding and documentation of the range of existing risks • Build/publish metrics, evaluation tools, methodological guidelines, protocols, and benchmarks for assessing risks of advanced AI • Develop/publish risk-based mitigation guidelines and safety mechanisms for advanced AI models, systems, and agents • Promote adoption of AISI guidelines, evaluations, and recommended AI safety measures and risk mitigations • Lead an inclusive, international network on the science of AI safety	• Develop and conduct evaluations on advanced AI systems • Drive foundational AI safety research • Facilitate information exchange	Wide range of functions including: • Implementing and enforcing EU AI Act; investigating breaches • Developing tools, methodologies, and benchmarks to evaluate capabilities • Monitoring the emergence of risks
Powers	No powers to make enforceable rules, compel evaluations, or access models or information without consent	No powers to make enforceable rules, compel evaluations, or access models or information without consent	Multiple powers, including to access documents, develop and conduct evaluations, take enforcement action
Collaborations	Intends to collaborate with US agencies, international partners, and diverse AI stakeholders. Has established AISI Consortium Has entered agreements with a number of large model providers to gain access to models for evaluations TTC established dialogue with with EU AI Office[25] MOU with UK AISI Network of AISIs	Entered agreements with large model providers to gain access to models for evaluations MOU with NIST AISI Network of AISIs	Mandate to cooperate with stakeholders from across sectors, other EU organs, and internationally[26] TTC established dialogue with with EU AI Office[27] Network of AISIs

In May 2024, the US, UK, and EU were joined by Australia, Canada, France, Germany, Italy, Japan, the Republic of Korea, and the Republic of Singapore in announcing the formation of a network of AISIs. According to the announcement:

> *Coming together, the network will build "complementarity and interoperability" between their technical work and approach to AI safety, to promote the safe, secure and trustworthy development of AI.*
>
> *This will include sharing information about models, their limitations, capabilities and risks, as well as monitoring specific "AI harms and safety incidents" where they occur and sharing resources to advance global understanding of the science around AI safety.*

The Network has the potential to advance interoperability in a number of ways. In particular, it is well-positioned to contribute to the development of the science of AI Safety and the development of tools, evaluations, metrics, and benchmarks for advanced foundation models.

The AISIs are tasked with conducting evaluations on advanced models. Fulfilling this function will inform both assessments of the capabilities and risks of particular models and inform safety research more generally. National AISIs are more likely to have access to models developed by domestic model providers.

The Network of AISIs provides a potential forum for information and knowledge about these models to be shared in a way that respects security and confidentiality concerns. The Network, therefore, has the potential to promote international collaboration on the science of AI Safety and the development of best practices that could feed into the development of interoperable policy frameworks at the national level.

Potential for future mutual recognition of evaluations

One interesting possibility arising from the creation of the AISI Network is the prospect of the **mutual recognition of evaluations conducted by national AISIs**. That could, in theory, promote interoperability at the institutional level, without requiring the development of fully aligned regulatory regimes. The Network has not yet been formally convened so it is difficult to assess how practical this may prove to be. Some factors that would support mutual recognition include:

- The AISIs will need appropriate access to models for evaluation. While the EU AI Office will have legal powers to access models and documentation, the UK and US AISIs are relying on a voluntary access model. It is too early to tell how this will affect the capacity of AISIs to examine models, though there have been reports that the UK AISI has not been able to access all models to date.
- More generally, the AISIs will need to have appropriate resources, functions, and powers to fulfill this mandate. National governments with AISIs participating in the network will need to be satisfied that other AISIs in the network are equipped to perform robust and reliable evaluations.
- An agreed basis for recognizing the assessments of national AISIs will be needed—including agreement on the testing and evaluations to be conducted, or the competency of other AISIs to develop/utilize appropriate methodologies to produce trustworthy assessments of model capabilities, risks, and mitigations. Agreement on what should be included in model documentation through establishing a good basis for best practice will also support mutual recognition efforts.

RECOMMENDATION 5A

National governments should continue to prioritize both international dialogue and collaboration on the science of AI Safety and the improved understanding of AI tools and models through initiatives such as the AI Summit series, the Interim International Scientific Report on the Safety of Advanced AI, and the UN International Scientific Panel on AI, however with more specificity and tracking of progress on commitments that will foster good practice.

RECOMMENDATION 5B

National governments should support the creation/development of AI Safety Institutes (or equivalent bodies), and ensure they have the resources, functions, and powers necessary to fulfill their core tasks (and in particular as a first focus, advancing the science of evaluation). Efforts should be made to align the functions of these bodies with those common among existing AISIs. (See Table 4 *above.)*

RECOMMENDATION 5C

The Network of AISIs (and bodies with equivalent or overlapping functions to existing AISIs such as the EU AI Office) should be supported and efforts should be made to expand its membership. Consideration should be given to how the Network could support broader AI Safety research initiatives—for instance, through sharing expertise gained by constituent AISIs in performing their functions, and inputting to other initiatives, such as the recently announced UN International Scientific Panel on AI.

Acknowledgments

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

We appreciate the invaluable input provided by experts who participated in consultations, provided written comments, or otherwise provided input into this paper, including:

- Rashad Abelson, OECD
- Anthony Barrett, UC Berkeley Center for Long-Term Cybersecurity
- William Bartholomew, Microsoft
- Alexandra Belias, Google Deepmind
- Dawn Bloxwich, Google Deepmind
- Miranda Bogen, Center for Democracy & Technology
- Kasia Chmielinski, Partnership on AI
- Amanda Craig, Microsoft
- Natasha Crampton, Microsoft
- Arisa Ema, University of Tokyo
- Evi Fuelle, Credo AI
- Lucia Gamboa, Credo AI
- Tiffany Georgievski, SonyAI
- Ani Gevorkian, Microsoft
- Alexandra Givens, Center for Democracy & Technology
- Sebastian Hallensleben, CEN/CENELEC
- William Isaac, Google Deepmind
- Antonia Kerle, BBC
- Cameron Kerry, Brookings Institution
- Ansgar Koene, EY
- Nada Madkour, UC Berkeley Center for Long-Term Cybersecurity
- Richard Mathenge, African Content Moderators Union
- Victoria Matthews, SonyAI
- Joshua Meltzer, Brookings Institution
- Valeria Milanes , Asociación por los Derechos Civiles
- Alondra Nelson, Institute for Advanced Study
- Marc Etienne Ouimette, AWS
- Lisa Pearlman, Apple
- Karine Perset, OECD
- Hadrien Pouget, Carnegie Endowment for International Peace
- Benjamin Prud'homme, Mila—Institut québécois d'intelligence artificielle
- Alejandro Segarra, Asociación por los Derechos Civiles
- Irene Solaiman, Hugging Face
- Andrew Strait, Ada Lovelace Institute
- Christian Troncoso, AWS
- Marcus Turner, A&O Shearman
- Raquel Vazquez Llorente, WITNESS
- David Wakeling, A&O Shearman
- Amy Winecoff, Center for Democracy and Technology
- Deon Woods Bell, Bill & Melinda Gates Foundation
- Harlan Yu, Upturn

We would also like to thank the Partnership on AI staff who contributed to this work, including Aimee Bataclan, Stephanie Bell, Rebecca Finlay, Jacob Pratt, Albert Tanjaya and Neil Uhl.

Endnotes

[1] National Telecommunications and Information Administration, Artificial Intelligence – Accountability Policy Report, March 2024, pp. 5, 37. Bengio, Yoshua et al., International Scientific Report on the Safety of Advanced AI - Interim Report. May 2024, p. 39.

[2] Bengio, Yoshua et al., International Scientific Report on the Safety of Advanced AI—Interim Report. May 2024, p. 37.

[3] UNESCO, Recommendation on the Ethics of Artificial Intelligence. November 23, 2021, p.22 [37].

[4] National Telecommunications and Information Administration, Artificial Intelligence—Accountability Policy Report, March 2024, pp 28ff; Chmielinski, K., et al., The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers. Shorenstein Center Discussion Paper, Mar 21, 2024, p18; Liang, Wiexin et al., "What's documented in AI? Systematic Analysis of 32K AI Model Cards". arXiv:2402.05160v1, February 7, 2024, p. 5.

[5] Chmielinski, K., et al., The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers. Shorenstein Center Discussion Paper, Mar 21, 2024, p. 14.

[6] NIST AI RMF Generative AI Profile, fn 1.

[7] EU AI Act, art 3(63).

[8] TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management. December 1, 2022.

[9] NIST AI RMF, p2

[10] Baroness Jones of Whitchurch, HL deb 30 July 2024, vol 839, col WA861, at https://hansard.parliament.uk/Lords/2024-07-30/debates/C1541E2E-0AE3-486C-9077-42CBA1785164/AITechnologyRegulations

[11] EU AI Act, art. 53.

[12] EU AI Act, art. 55.

[13] EU AI Act, art. 51(1).

[14] EU AI Act, art. 3.

[15] EU AI Act, Annex XIII.

[16] EU AI Act, art. 3(65).

[17] AI Executive Order, sec 2(k) (emphasis added).

[18] AI Executive Order, s. 4.2(b)(i).

[19] AI Executive Order, sec. 4.2(b).

[20] EU AI Act, arts 53(4), 55(2).

[21] Kerry, Cameron F., Small Yards, Big Tents: How to Build Cooperation on Critical International Standards. Brookings Institution Report, March 2024, p. 2.

[22] NIST, The United States Artificial Intelligence Safety Institute: Vision, Mission, and Strategic Goals, May 21, 2024.

[23] DSIT, Introducing the AI Safety Institute. UK Government Policy Paper, updated January 17, 2024.

[24] "European AI Office", Webpage, updated August 1, 2024.

[25] Joint Statement EU-US Trade and Technology Council of 4-5 April 2024. April 5, 2024.

[26] Commission Decision Establishing the European AI Office. January 24, 2024.

[27] Joint Statement EU-US Trade and Technology Council of 4-5 April 2024. April 5, 2024.