PARTNERSHIP ON AI

Partnership on AI's Response to National Telecommunications and Information Administration—AI Accountability Policy Request for Comment

Docket NTIA-230407-0093

Partnership on AI (PAI) is a non-profit partnership of academic, civil society, industry, and media organizations creating solutions to ensure that AI advances positive outcomes for people and society. PAI studies and formulates sociotechnical approaches aimed at achieving the responsible development of artificial intelligence (AI) and machine learning (ML) technologies. Today, we connect over 100 partner organizations in 14 countries to be a uniting force for the responsible development and fielding of AI technologies.

PAI develops tools, recommendations, and other resources by inviting multi-stakeholder voices from across the AI community and beyond to share insights that can be synthesized into actionable guidance. We then work to promote adoption in practice, inform public policy, and advance public understanding. We are not an industry or trade group nor an advocacy organization. We aim to change practice, inform policy, and advance understanding.

The information in this document is provided by PAI and is not intended to reflect the view of any particular Partner organization of PAI. The comments provided herein are intended to provide evidence-based information, based on PAI's research, in response to the NTIA's questions.

Overview of Partnership on AI's Recommendations

- Accountability measures should be considered across the AI Value Chain and **different accountability mechanisms might be needed for different stakeholders depending on their roles**. For example, a proportion of the transparency mechanisms and steps targeted at those (a) developing, (b) creating, or (c) distributing synthetic media will need to be different, and tied to their unique tasks and responsibilities.
- **Accountability needs to be informed by different perspectives.** Civil Society, Academia (e.g., for red-teaming and evaluation), and the public (including the voices of workers) distinctly play a role.
- **There will be accountability measures which apply across different use cases** (e.g., an impact assessment for the impact on workers), and those which will need to be **targeted at specific risks related to applications/cases**.
- **Transparency is a key lever for achieving accountability.** PAI advocates for a "full lifecycle" approach, meaning that policy related to documentation should start before the first phase of development.
- **Efforts to drive multistakeholder dialogue and solutions are critical in the absence of consensus on the most important risks related to frontier AI models and how they should be mitigated.** There is a strong need for convening and the NTIA should support the development of protocols and dialogue which seek to drive alignment. PAI has begun the coalition-building with our multistakeholder dialogue to develop shared protocols for responsible large-scale model deployment.

This submission is organized around three topics

In response to the National Telecommunications and Information Administration's (NTIA) request for comment on AI accountability policy, Partnership on AI provides a short overview of best practices focused on:

1. Accountability across the AI Value Chain with different responsibilities for different stakeholders
2. Transparency tools as a way to support accountability
3. Accountability and Foundation models

1. Accountability across the AI Value Chain: Responsibilities for different stakeholders

The AI Value Chain (or supply chain) is complex, often involving open-source and proprietary products and downstream applications that are quite different from what AI system developers may initially have contemplated. Moreover, training data for AI systems may be acquired from multiple sources, including from the customer using the technology. Problems in AI systems may arise downstream at the deployment or customization stage or upstream during model development and data training.

Accountability must be integrated across the AI pipeline, spanning AI development, deployment, and maintenance and carried out by the institutional and individual stakeholders contributing to the AI supply chain. Values like transparency and disclosure are consistent across stakeholders, however, the ways in which different stakeholders can implement these values promoting accountability can vary.

Accountability needs to be shaped and informed by all affected stakeholders to account for not just technical experts in the space but also experiential experts. It is important that policymakers identify the key stakeholders of each sector and understand what information they require to have literacy of the systems that affect them. The first step is to be able to meaningfully and ethically engage with such stakeholders. PAI's white paper Ethical Principles and Practices for Inclusive AI outlines four guiding principles for ethical engagement grounded in best practices:

1. All participation is a form of labor that should be recognized
2. Stakeholder engagement must address inherent power asymmetries
3. Inclusion and participation can be integrated across all stages of the development lifecycle
4. Inclusion and participation must be integrated into the application of other responsible AI principles

Designing policy frameworks to account for the roles of different stakeholders from development to deployment will be important. Below, we share case examples that are testing for and requesting specific action to mitigate harms linked to a specific form of application/type of output (e.g., generative media) or to assess impact in relation to a specific group (e.g., on workers).

CASE EXAMPLE 1

Designing accountability measures for different stakeholders across the AI Value Chain: Synthetic Media Framework

Partnership on AI's Responsible Practices for Synthetic Media (Framework), a normative technology framework on generative (or synthetic) media, targets the pipeline of AI development and deployment, with specific actions for:

1. Entities building AI technology and infrastructure
2. AI creators
3. AI distributors, which includes both technology platforms and media institutions

Technology builders and developers, for example, are encouraged to "provide disclosure mechanisms for those creating and distributing synthetic media" and build out tools like media provenance infrastructure. Creators and distributors are given different guidance about the use of these available tools for labeling content and ensuring accountability.

PAI offers recommendations for different categories of stakeholders with regard to their roles in developing, creating, and distributing synthetic media.[1] However, it is important to note that these categories are not mutually exclusive. A given stakeholder could fit within several categories, as in the case of social media platforms.

[1] Here, synthetic media, also referred to as generative media, is defined as visual, auditory, or multimodal content that has been generated or modified (commonly via artificial intelligence). Such outputs are often highly realistic, would not be identifiable as synthetic to the average person, and may simulate artifacts, persons, or events.

CASE EXAMPLE 2

Guidelines for AI and Shared Prosperity

PAI has recently released the Guidelines for AI and Shared Prosperity: a set of tools to inform the design and development of workplace AI systems to protect workers' rights and well-being. The Guidelines center the insights from frontline workers (published in PAI's AI and Job Quality report) and were developed under the guidance of a multidisciplinary Steering Committee, consisting of senior leaders from the labor movement, the technology industry, civil society, and academia. The Guidelines address the impacts of AI on the broader labor market and contain additional guidance on creating acceptable working conditions for workers who contribute to dataset creation and enrichment: critical inputs to the AI Value Chain.

The Guidelines offer two tools:

1. **A high-level Job Impact Assessment Tool**
2. **A collection of Responsible Practices and Suggested Uses** tailored for specific stakeholder groups:
   - AI-developing organizations
   - AI-using organizations
   - Labor organizations
   - Policymakers

The Guidelines serve as an accountability mechanism for those organizations. They provide AI-developing and AI-using organizations with tools to hold themselves accountable to the people their decisions affect, and equip workers and their representatives to ensure harmful decisions are not made at their expense. Through their use, risks to workers and the broader economy from AI design and deployment can be identified early and effectively mitigated, while opportunities to improve job access and job quality are maximized.

2. Transparency as a driver of AI Accountability

PAI's ABOUT ML (Annotation and Benchmarking on Understanding and Transparency of Machine Learning) workstream answers the question, "If organizations and companies value the principles of fairness, transparency, and accountability, how might we operationalize those principles?"

The ABOUT ML work advocates for the transparency of machine learning systems through documentation of the machine learning lifecycle, and **we recommend that the NTIA builds documentation into any accountability and evaluation model, to ensure documentation is utilized across the ML lifecycle.** PAI's ABOUT ML research program has gained insight from world leading experts over the years. Their insight informs this recommendation.

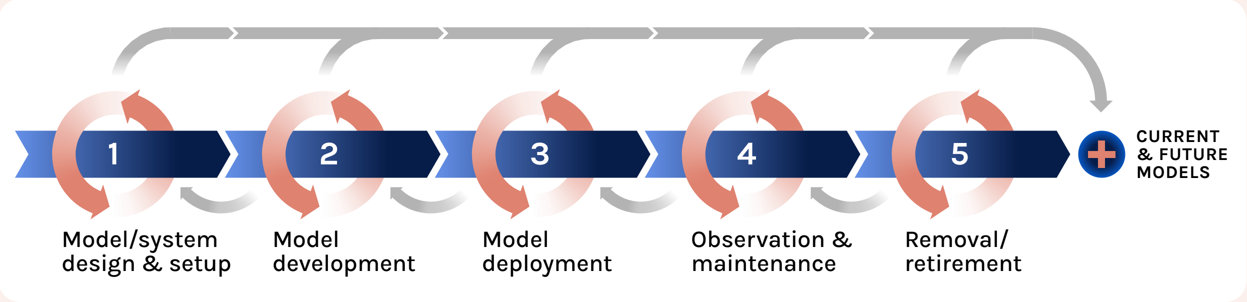
It is not simply about disclosing a list of characteristics about the data sets and mathematical models within an ML system, but rather an entire process that an organization needs to incorporate throughout the design, development, and deployment of the ML system being considered. In this section, we provide the NTIA with guidance on documentation as a method to support accountability, with a particular focus on documentation and answer questions such as:

- What sorts of records (e.g., logs, versions, model selection, data selection) and other documentation should developers and deployers of AI systems keep in order to support AI accountability?
- How long should this documentation be retained? Are there design principles (including technical design) for AI systems that would foster accountability by design?

Best practices for documentation as a method for growing transparency and accountability

1. **Documentation practices cannot remain within the sole purview of practitioners.** In a qualitative study with practitioners, PAI found that a major challenge in implementing documentation standards within organizations was shifting organizational practice and culture to prioritize documentation against other tasks. Guidance or regulatory requirements may incentivize practitioners to improve their documentation practices, not only by requiring the publication of dataset and model provenance but by reiterating the public value of such documentation.
2. **PAI advocates for a "full lifecycle" approach to considering transparency and accountability mechanisms.** *(See illustration below.)* We propose that it starts before phase one, "Model System Design and Setup," and considers questions about data specification, curation, and integration as well as model specification. Robust documentation involves updates to existing documentation and the creation of different documentation that track through "Model Training and Evaluation," "Model Deployment," and finally through "Model and Data Maintenance, Further Integration or Retirement."

A Guide For Documentation:
Steps to Integrate into Policy Design



The considerations identified here serve as the bare minimum of how each stage of the ML lifecycle should be considered and evaluated. The full context surrounding various data and model steps should be integrated into any policy designed for documentation.

For a comprehensive list of questions and considerations please see the ABOUT ML Process Guide.

Stage 1 Documentation: Model System Design and Setup

Data Specification	Documenting the motivation for producing a dataset provides a lever for accountability as the project proceeds, enabling stakeholders to go back to the original intent behind a dataset's creation and check that the current trajectory tracks with the original goal.
Data Curation	Collection, processing, composition, and various judgment calls (often made when creating, correcting, annotating, striking out, weighing, or enriching data) are all parts of the curation process. A key area to document is how you are using your test, training, and validation datasets.
Data Integration	Data integration might include connecting data sources, data distribution to users, the inclusion of data pipeline and data management, maintenance of data, and audits of data usage and issues. It might be useful to separate the technical and data science aspect from the safety and continuous use of data.
Model Specification	Key questions to document include the choice of structure, choice of output structure, choice of loss function and regularization, where random seeds come from and where they are saved, hyperparameters, optimization algorithm, and generalizability measured by how much difference in a test they expect to see

Stage 2 Documentation: Model System Training and Evaluation

Model Training	It is important to share how the model was architected and trained, and the process that was used for debugging.
	Choices of ML model architectures have numerous consequences that are relevant to downstream users, so it is essential to document both the choices and the rationales behind them.
Model Evaluation	It is important to help users of the model understand how the model was checked for accuracy, bias, and other important properties.
	The specific metrics that the model will be tested on depend on the particular use case, so it is helpful for this documentation to include examples of which metrics apply for which use case.

Stage 3 Documentation: Model System Deployment

Model Integration	Even if each portion of the model is thoroughly tested, validated, and documented, there are additional documentation and evaluation needs that arise when connecting a model into a broader ML system. Validating how the models interoperate is important because models might not work well together.

Stage 4 Documentation: Model System Data and Model Maintenance

Data Maintenance	This is important for helping users know whether they are using the latest dataset and whether the dataset will be kept up-to-date. If the dataset is not maintained, however, there can be concerns about the interpretability and applicability of the dataset.
Model Maintenance	There are many parameters of a model update which would be useful to document: how and why an update is triggered, whether old models or parameters can still be accessed, who owns maintenance and updating, and guidance on reasonable shelf life of the model, including performance expectations for versions.

CASE EXAMPLE 3

Interrogating the Risks and Benefits of Demographic Data Collection, Use, and Non-Use

PAI's research on AI accountability goals (including this report, blog post, and conference paper) highlights a number of trade-offs related to assessing the social harms of AI systems (e.g., algorithmic discrimination, bias, and unfairness that impact individual livelihoods and well-being). **In order for organizations to assess whether or not their system is discriminatory, they need to have access to demographic data in order to make performance comparisons or standardizations across groups.**

However, this data is rarely accessible in practice, due to a range of legal concerns (such as anti-discrimination law and policies in the US and privacy policies) and organizational barriers (such as concern over reputational harm and misalignment with organizational goals).

This leaves practitioners at a loss for how to adequately assess for discrimination in their systems. Choosing not to collect demographic data to assess for discrimination means that discrimination will remain invisible and likely unaddressed.

For organizations seeking to collect demographic data to support assessments, there are important trade-offs to consider when working towards mitigating discrimination, including: violating personal privacy, misrepresenting individuals, using sensitive data beyond what was consented to, increasing surveillance of disenfranchised groups, reinforcing oppressive or overly prescriptive categories, and control by private entities over what constitutes bias or harmful treatment (as opposed to how the data subject's define harm for themselves). **By confronting these questions before and during the collection of demographic data, algorithmic fairness methods are more likely to actually mitigate harmful treatment disparities without reinforcing systems of oppression.**

3. Accountability and Foundation Models

The rapid pace of large-scale AI deployment and its potential to negatively impact communities at scale require us to act now, aligning on best practices based on shared insights. In the absence of consensus on the most important risks and how they should be mitigated, we believe this alignment requires multistakeholder input from a diverse set of perspectives. In addition to the consideration of regulatory interventions, non-regulatory approaches to responsible deployment — such as voluntary norms or protocols — are important to consider as complimentary levers given that the full downstream impacts of this technology are not yet known.

Pre-deployment steps are critical. As the technical foundation of many consumer-facing AI applications (and those to come), large-scale models and how they are released can have an outsized impact on end users. This makes the steps in the path from development to deployment (which may include pre-deployment testing, risk identification and mitigation, monitoring, and oversight, including the possibility of halting deployment) particularly important to consider.

We must also address the unique and novel nature of the latest AI models and systems that are to come. This means conducting:

- Risk assessments
- Internal and external audits
- Red-teaming to identify emerging risks and security vulnerabilities, monitoring systems after deployment

Additionally, an ecosystem of third party actors should be cultivated to enable external access and scrutiny of these systems.

Multistakeholder dialogue should be a core part of NTIA's strategy. Partnership on AI has begun leading an effort to address both a shared understanding of long-term risks as well as building a shared consensus of what accountability protocols might look like. We have already kicked-off a multistakeholder dialogue to develop shared protocols for responsible large-scale model deployment. PAI believes that collaboration between actors with diverse perspectives is crucial to fully understanding risks related to large-scale AI. To facilitate our multistakeholder approach, this work has been guided by PAI's Safety Critical AI steering committee. Made up of experts from the Alan Turing Institute, the American Civil Liberties Union, Anthropic, DeepMind, IBM, Meta, and the Schwartz Reisman Institute for Technology and Society, among other organizations, this recently formed steering committee has convened over the last few months to do the vital work of identifying concrete interventions where community collaboration can make the greatest impact. **We welcome NTIA's engagement in this work as we take it forward.**