Hugging Face
12 June 2023

20 Jay St
Suite 620
New York, NY 11201

Hugging Face Comments on AI Accountability for the National Telecommunications and Information Administration

Hugging Face commends the National Telecommunications and Information Administration (NTIA) Task Force on its extensive work framing the different aspects and components of accountability of AI systems. The following comments are informed by our experiences as an open platform for state-of-the-art (SotA) AI systems, working to make AI accessible and broadly available to researchers for responsible development. Comments are organized by section and by question. If a section or question is not highlighted, we do not have specific, actionable feedback.

About Hugging Face

Hugging Face is a community-oriented company based in the U.S. and France working to democratize good Machine Learning (ML), and has become the most widely used platform for sharing and collaborating on ML systems. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analyzing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI.

AI Accountability Objectives

Question 1: What is the purpose of AI accountability mechanisms such as certifications, audits, and assessments?

As new applications of technology move from academic research and experimental development to broader integration and deployment within society, they start shaping the lives and experiences of millions to billions of direct and indirect users. Accountability mechanisms such as certifications, audits, or assessments help ensure that the safety, security, and well-being of these users are formally taken into account, and that the many choices that frame the technology's development verifiably contribute to addressing the potential negative societal impacts that may come from insufficient foresight or consideration.

These accountability mechanisms are particularly needed for AI-enabled technology — especially Machine Learning (ML) systems — on two main accounts. First, as a data-driven technology, ML constitutes a sociotechnical system. The many modes of interaction between data subjects, algorithm subjects, and the various stages of the technology development make it particularly difficult to predict the impact of a single development choice in isolation, and make the technology prone to exacerbating social biases and historical discrimination [1]. The severity of these risks is further increased by AI systems' ability to be deployed at scales that were previously harder to achieve.

Second, Machine Learning is shifting from an academic discipline whose benchmarks were developed to compare the contribution of different scientific innovations in an open and controlled setting to a product race in a multi-billion dollar market with very different constraints and priorities. This shift is creating an accountability gap, one that is particularly marked with respect to the technology's social impact. Indeed:

1. The academic benchmarks that have shaped the development of ML technology to date have targeted mostly internally relevant notions of performance and generalization, to the detriment of evaluation of in-context behavior and potential negative impacts [2]. Even evaluations that do address phenomena such as bias and discrimination tend to do so in an abstract way that tends to lack proper contextualization, limiting their utility in preventing real-world harms when systems are deployed [3].
2. **Performance on an academic benchmark is not by itself a sufficient guarantee of appropriateness for real-world deployment.** Adopting ML systems for a range of applications without demonstrating their fitness for the purpose has led to a crisis referred to as "Fallacy of A functionality" [4] and prompted the FTC to remind developers to better align their claims about the positive contributions of their AI systems with actual verifiable and measurable improvements [5].
3. Shifting evaluation of ML systems from an open setting that allows for external scrutiny and reproduction to evaluation behind closed doors without mechanisms for independent verification is also putting their reliability and legitimacy into question - and in particular raising questions about data contamination, a common issue with ML system evaluation [6][7]. In particular, **transparency has proven necessary to evaluate how ML artifacts may exacerbate risks of discrimination** [8], whereas interventions that aim to address these issues in commercial systems without external accountability have proven unreliable in the past [9].
4. Finally, as systems are deployed at scale, the **introduction of AI may shift burdens in a way that benefits the deployer but incurs additional costs for other parts of society,** as has been shown by the additional work required of education professionals over the last year in the wake of the deployment of generative AI as a product [10].

Given both the nature of these systems and the rapid shift in their evaluation needs, new accountability mechanisms are thus required and should focus on defining and verifying due diligence in terms of:

1. **Verifying that a marketed or deployed system is evaluated for functionality in a well-defined application case.** This includes limiting the cost incurred by society at large when an entity deploys a system at scale that performs differently from what has been advertised

2. **Mitigating ML's risk of exacerbating systemic issues,** such as historical discrimination
3. **Providing basic security guarantees to users** of the system, including but not limited to ensuring a safe workplace, promoting information security, and putting safeguards so that unintentional potentially harmful side-effects of the technology application are detected and mitigated in a timely manner.

Question 2: Is the value of certifications, audits, and assessments mostly to promote trust for external stakeholders or is it to change internal processes? How might the answer influence policy design?

Accountability mechanisms serve a dual role of:

1. shaping the development and deployment of AI-enabled technology, and

2. providing assurances that broadly used technology meets a minimum standard of reliability, responsibility, and regulatory compliance.

On (1), the technology development and deployment is shaped by the requirement to think through, learn about, and address the different questions or specifications relevant to the accountability mechanism. It gives developers a concrete specification of their responsibilities in terms of the broader impact of their technology, one that is subject to governance by the appropriate public and governmental institutions and that they shall prioritize on the same level as other performance targets; in other words, they codify a notion of due diligence and social responsibility of the entities putting AI systems into service. Additionally, **when accountability mechanisms are paired with transparency (such as open reporting of results to a third party), "good practices" are incentivized** by virtue of the fact that technologists will generally not want to show poor work or bad results.

On (2), **certifications, audits, assessments help close the accountability gap** introduced by the shift from science to product by verifying claims. They also encourage good practice by making sure that the systems are ready to show proof of good practice and verification.

AI accountability mechanisms that leverage transparency requirements also promote a healthier ecosystem of open research and collaborative development, by letting all stakeholders make informed comments and contributions to how the technology should be shaped — including the direct and indirect users who best know how their needs should shape the technology. With the right compliance support for small and medium actors, accountability mechanisms will speed up, rather than slow down, the development of technology for the benefit of all.

Both **internal and external audits have complementary roles** to play. Internal audits can surface the necessary information to pre-certify systems and guide developers in their effort to examine their own work. They are particularly useful in cases that require monitoring of a system behavior over a lengthy period of time, such as monitoring failures, evaluating performance drift, and providing early warning of discriminatory outcome trends. External audits on the other hand can be warranted either to empower external stakeholders who have reason to suspect the system is engendering harms, or to verify the faithfulness and accuracy of documentation produced by the system developer [11].

With these considerations in mind, we recommend the following:

- In order to promote good practice in technology development, **internal accountability mechanisms should focus on process as well as measures and metrics** in defining obligations for the developers.
- Both internal and external accountability mechanisms should **prioritize transparency** as well as sharing results and documentation as often as possible, to allow for more diverse contributions and more informed choices from potential users and affected stakeholders.
- External accountability mechanisms should prioritize agency by and communication to external stakeholders. External accountability processes can respond to concerns about the behavior of a system, and should include provisions for responding to requests from affected populations. They should document their findings in a way that is primarily intelligible to those categories of stakeholders.

Question 5: Given the likely integration of generative AI tools such as large language models (e.g., ChatGPT) or other general-purpose AI or foundational models into downstream products, how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI?

General-purpose AI systems (GPAIs) outline the strong need for modular and properly articulated accountability mechanisms at every stage of the development chain, as well as the importance of transparency as a general requirement. GPAIs constitute a class of ML systems that can support a broad range of applications - in some cases because they can respond to varied forms of inputs, and often with some additional work to adapt them to a specific use case through techniques like Parameter-Efficient Fine-Tuning (PEFT)[33][34] that allow for fine-tuning at a fraction of the original training cost.

GPAIs present a unique opportunity to develop AI-enabled technology much easier than if developers had to start from scratch for every application. They also present unique challenges for accountability, and exacerbate all of the issues outlined above (see our answer to Question 1). In particular, they are often marketed as systems that can be applied to *any* settings without sufficient validation [5]. They also typically introduce more distance between their initial development setting and their practical application in technology, making it harder for down-stream users to understand *e.g.* how the initial training data choices may give rise to risks of discrimination or other harms in a specific use case [1].

Extensive literature has outlined the role of transparency and documentation in addressing these challenges. **Dataset Sheets [12] and Data Statements [13] provide an entry point into**

the datasets that shape the GPAIs, have seen broad adoption, and have inspired the Dataset Cards used to describe datasets on the Hugging Face Hub [14]. When accompanied by additional tooling and visualization of the content of a dataset [15][16], they can help support *post-hoc* analysis of a system's fitness for a specific use case. **Model cards [17] play a similar role in giving a model's prospective user a summary of the model's main characteristics**, including its performance on established benchmarks, its biases and limitations, and uses that may be intended by the developers or, conversely, uses that may be out of scope [18]. **New licensing schemes such as Responsible AI Licenses [19] also help provide a legal framework** for this specification by letting developers of GPAIs make the systems available for broad uses while prohibiting applications that are particularly likely to cause harm without proper additional testing and guardrails [20].

Documentation belongs to the category of "internal" accountability mechanisms. While those are necessary to promote responsible technology development, **they are not by themselves sufficient** for trustworthy and reliable technology [11]. Even when entities do their best to document biases in their systems, there is only so much a small team with mostly technical expertise can analyze — especially when the teams' primary focus is on other definitions of technical performance that are perceived to be more instrumental to the product's success. In practice, **external scrutiny of commercial systems** has been instrumental in showcasing how their biases may directly affect downstream users, from face recognition [21] to image generation [24] and chatbots [23].

The **method of release of GPAIs affects what types of safeguards are more applicable** to a given system; and consequently what accountability mechanisms are most relevant. For example, an AI model deployed via an API can be rate-limited, where users are limited to a certain number of outputs per timeframe to prevent mass generation and harms such as disinformation spreading. Ultimately, safeguards cannot be solely technical and should rely on a combination of measures, such as robust documentation and community feedback. Mechanisms for trustworthiness can also be applied at the software, hardware, and institutional levels.

Question 6: The application of accountability measures (whether voluntary or regulatory) is more straightforward for some trustworthy AI goals than for others. With respect to which trustworthy AI goals are there existing requirements or standards? Are there any trustworthy AI goals that are not amenable to requirements or standards? How should accountability policies, whether governmental or non-governmental, treat these differences?

Safety and effectiveness are most amenable to traditional requirements and standards. Notice and explanation and human alternatives require appropriate scoping, but are easier to verify.

Data privacy stands out as a trustworthy AI goal that is strongly process-based, and depends chiefly on good data governance and on the formalization and enforcement of individual privacy rights [25]. While the governance requirements may depend on some technical aspects of the

ML systems, such as their likelihood of memorizing individual data points [26], **the most efficient privacy protections come from the data selection, processing, and management** [27].

Algorithmic discrimination comes from the combination of the biases encoded in the ML systems and the choice of deployment settings and application [28]. While strong requirements and standards at both levels are necessary to lowering the likelihood of discrimination, neither should be thought to be sufficient. At the system level, **model developers should provide sufficient information so that a model's inherent biases can be evaluated and re-evaluated in the context of new applications,** whether they commit to working on these evaluations themselves or provide sufficient access to the model and dataset for other entities to do so. These evaluations should focus on surfacing patterns in how protected categories are represented in the data and model. At the deployment level, developers should be held accountable for choosing ML systems that have the least risk of perpetuating historical discriminations in their application setting. Regulatory tools like the EU AI Liability directive [29] can help clarify these requirements and the developers' responsibility for negative outcomes in a particular application setting.

Question 7: Are there ways in which accountability mechanisms are unlikely to further, and might even frustrate, the development of trustworthy AI? Are there accountability mechanisms that unduly impact AI innovation and the competitiveness of U.S. developers?

Given the breadth of the new applications of AI technology and the scale of its deployment, accountability mechanisms need to be able to leverage inputs from as much of society as possible, including developers, academic researchers, advocacy groups, and journalists to support regulators and agencies in their goals. While the companies and individuals developing the technology should be encouraged to do their best to take the safety of their systems into account, requiring them to understand the full scope of social impact of a technology with such rapid development and adoption as well as navigate complex value tension before different stakeholders is both unrealistic and contrary to democratic values.

To that end, these mechanisms should favor transparency whenever possible, so that said stakeholders may interrogate and critique the development choices before they lead to significant harms. They should also consider the needs of nonprofit actors, academic researchers, and small and medium enterprises who can explore alternative ways of building ML systems in open settings, by facilitating regulatory compliance for less-resourced entities. Contrary to common narratives, regulation that drives a greater range of actors to innovate new ways of developing robust and reliable technology accelerates a holistic definition of progress, and makes US companies more competitive in anything but the shortest time horizon.

Conversely, **we caution against accountability mechanisms that are likely to take governance of AI systems out of public awareness and enshrine the role of a few**

companies in unilaterally shaping its development - at a risk of giving them an outsized influence on the values embedded in this technology.

Accountability Subjects

Question 15: The AI value 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.

a. Where in the value chain should accountability efforts focus?

Accountability efforts need to be distributed across the value chain, since most harms enabled by AI systems come from a combination of decisions made at various levels of the development process.

Training dataset collection and management should respect data subject rights, including privacy and applicable intellectual property rights. Training data selection encodes specific values and behaviors, including harmful social biases, into ML systems, which should be the focus of accountability mechanisms for all actors downstream in the development chain. Model training algorithms have a direct influence on what the model memorizes and encodes. Controls at the deployment stage can monitor for unwanted behaviors and help mitigate risks by triggering timely interventions. Finally, choosing where and when to deploy an AI system, and which natural persons will be actively and passively affected by it, shapes both who benefits and who bears the risk from AI deployment.

While it could be tempting to focus accountability efforts purely on the last stage of deployment, such an approach risks introducing a mis-alignment between the work and forethought required at each stage to promote responsible development and the capacity of the different actors across the development chain. **Explicitly distributing accountability makes it more likely that each actor will have better tools and ML components to work with, and be better able to meet their own requirements;** especially in collaborative development settings including academic and grassroots research organizations.

b. How can accountability efforts at different points in the value chain best be coordinated and communicated?

Given the dependence of accountability efforts at any point of the value chain on decisions made at other points, it is paramount that people working both on the development and on audits or assessments have access to sufficient information about development choices and about the outcomes of assessments by other actors.

This access can take several forms. Access to technical artifacts and documentation (including documentation created by audits and assessment) can be granted to the general public, to researchers, developers, and auditors who work directly with the artifacts, or solely to accredited external auditors. Projects like BigScience [30] and BigCode [31] showcase an approach for combining the first two options. In both projects, datasets and models prioritized fully open release whenever possible, and provide access on a case-by-case basis for datasets or artifacts with stronger privacy concerns [32]. For data, specifically, when full access cannot be provided to external researchers, a minimum standard of disclosure should cover **when the data was collected, from what sources, and how it was processed**.

Whatever the level of access to the artifacts, strong standards for documentation, including minimal requirements for the information that needs to be included in a dataset or model card, can help actors across the value chain make informed decisions that can then be held accountable for their consideration of their social impacts. In particular, the **level of information provided should be sufficient to support the analysis required when a component is used in a new application setting**, particularly in the case of components that are used in the development of General Purpose AI systems.

Question 16: The lifecycle of any given AI system or component also presents distinct junctures for assessment, audit, and other measures. For example, in the case of bias, it has been shown that “[b]ias is prevalent in the assumptions about which data should be used, what AI models should be developed, where the AI system should be placed—or if AI is required at all.” How should AI accountability mechanisms consider the AI lifecycle?

a. Should AI accountability mechanisms focus narrowly on the technical characteristics of a defined model and relevant data? Or should they feature other aspects of the socio- technical system, including the system in which the AI is embedded? When is the narrower scope better and when is the broader better? How can the scope and limitations of the accountability mechanism be effectively communicated to outside stakeholders?

Both aspects of accountability mechanisms are necessary and complementary. Some aspects of a base system or dataset can and should be measured in isolation, including its performance on technical benchmarks and some categories of encoded biases. Some aspects of accountability mechanisms should focus on the impact of a system in use, including risk monitoring, robustness to changes in the input distribution, and fairness at the level of the impact on the active and passive users (as opposed to at the level of the model outputs). A standardized label for ML systems that outlines what aspects have or have not been assessed as well as the outcome of the assessment can help stakeholders better contextualize the behavior of AI systems, and trigger further investigation as required.

b. How should AI audits or assessments be timed? At what stage of design, development, and deployment should they take place to provide meaningful accountability?

AI assessments of ML components such as datasets or models can be most effective at the following stages:

- **When a component is first made available**, either in limited capacity to actors working on a different part of the development chain or when it is shared more broadly:
    - For example: a report detailing training data, pre-processing, model size, and in-scope vs out-of-scope applications.
- **When a component is integrated in a system**, or used in a different part of the development chain:
    - Biases and performance can be evaluated in the context of the new proposed use. For example, the first time a language dataset is used to pretrain a system to support automatic content moderation can trigger new analysis of the sentiment associated with different marginalized identities in the dataset.
- **When an AI system using the component is deployed in a concrete application setting:**
    - A well scoped application provides further information about which aspects of a model or dataset should be the focus of additional scrutiny.

In all of the cases outlined above, the **results of the assessment should be added to the documentation of the component,** allowing users and researchers to more easily benefit from this work.

Accountability Inputs and Transparency

Question 22: How should the accountability process address data quality and data voids of different kinds? For example, in the context of automated employment decision tools, there may be no historical data available for assessing the performance of a newly deployed, custom-built tool. For a tool deployed by other firms, there may be data a vendor has access to, but the audited firm itself lacks. In some cases, the vendor itself may have intentionally limited its own data collection and access for privacy and security purposes. How should AI accountability requirements or practices deal with these data issues? What should be the roles of government, civil society, and academia in providing useful data sets (synthetic or otherwise) to fill gaps and create equitable access to data?

Curating appropriate test sets should be a part of the development process for deployed systems; if a model has not been evaluated across performance and risk considerations, it should not be commercially deployed. In order to facilitate this accountability process, and as part of its implementation, companies should be incentivized to make as much of these test sets available as possible — using pseudonymization or other post-processing as necessary to protect the privacy of their users. Maintaining and growing such a commons of in-context performance and bias evaluations will help significantly advance our understanding of risks tied to concerns such as representational harms. Additionally, agencies responsible for conducting audits can help bootstrap evaluation benchmarks for novel tasks and help maintain them with contributions from audited entities.

Contracts between vendors and audited entities should include limited additional data transfer to the auditing entity in case of audits. Agencies may help vendors identify what subsets of the data are relevant, and use it for the sole purpose of the audit.

Question 23: How should AI accountability "products" (*e.g.,* audit results) be communicated to different stakeholders? Should there be standardized reporting within a sector and/or across sectors? How should the translational work of communicating AI accountability results to affected people and communities be done and supported?

We strongly advocate for **a centralized repository of audits, which will help build stronger methodology and avoid duplication of work** (e.g. by re-evaluating common components of systems). We recommend the repository be as open as possible, which will allow civil society and advocacy organizations to participate in the translational work of communicating and interpreting the results.

Standardized records should describe the targets of the audit, the methodology, and the tools used. Components of the analysis should be directly shared when possible, extensively described when not. A centralized repository will help shape good practices and a common understanding of what constitutes sufficient documentation of the process between different stakeholders.

Barriers to Effective Accountability

Question 24: What are the most significant barriers to effective AI accountability in the private sector, including barriers to independent AI audits, whether cooperative or adversarial? What are the best strategies and interventions to overcome these barriers?

The largest barriers are the lack of information shared about systems, lack of legal protections for novel AI impacts, the complexity of impact evaluations, and the stochastic nature of AI system behaviors. System components include models and datasets, but also filters and additional technical and decision-making documentation. Furthermore, there is currently no legal mandate that users are informed about their interaction with an AI system. **Mandatory disclosure of an AI system's use and labeling content as AI generated should include information uniquely identifying the AI system used.**

In addition, current legal mechanisms create barriers for individuals to gain information about systems; Freedom of Information Act (FOIA) requests are the closest equivalent but are hard to enforce and operationalize without experience. **Legal mechanisms should be created for affected individuals to inspect specific system decisions.**

The stochastic nature of the harms can make them unreliable to reproduce and investigate. Developers' are often able to patch for specific cases as they are reported without addressing underlying issues, further complicating the process. This issue could be partly addressed by requiring entities deploying models to clearly mark changes in the versions of the underlying models, and ideally by letting users opt to access previous versions for the purpose of investigating specific behaviors.

Finally, evaluating systems for their fitness for purpose requires robust evaluations across functions that are not standardized and differ by system type. Research toward evaluating social impacts of systems is ongoing, but sees large gaps in assessing or quantifying inherently qualitative aspects of outputs.

Question 25: Is the lack of a general federal data protection or privacy law a barrier to effective AI accountability?

Comprehensive privacy legislation, such as the EU's General Data Protection Regulation (GDPR), is an important instrument to govern the use of personal data by AI systems. **Federal data protection would give a direct recourse to data subjects and mandate responsible data management.**

Question 26: Is the lack of a federal law focused on AI systems a barrier to effective AI accountability?

Federal regulation should evolve in light of AI systems, which is different from having a single piece of legislation on AI systems. Updated definitions should appear in relevant regulations. Better definitions are needed for: personal data; non-discrimination and liability; due diligence in evaluating application-specific AI risks; and right to explanation in sectoral regulations.

Question 27: What is the role of intellectual property rights, terms of service, contractual obligations, or other legal entitlements in fostering or impeding a robust AI accountability ecosystem? For example, do nondisclosure agreements or trade secret protections impede the assessment or audit of AI systems and processes? If so, what legal or policy developments are needed to ensure an effective accountability framework?

Fostering novel legal approaches such as RAIL licenses and similar terms of service, if sufficiently well scoped, can help support the development of generally useful systems while holding users accountable across the value chain. Trade secrets and nondisclosure agreements cannot override the need to share technical details for the purpose of external auditing.

Question 28: What do AI audits and assessments cost? Which entities should be expected to bear these costs? What are the possible consequences of AI accountability requirements that might impose significant costs on regulated entities? Are there ways to reduce these costs? What are the best ways to consider costs in relation to benefits?

AI audits and assessments incur different categories of costs, including:

- technical expertise - an audit requires time from a person with the expertise to understand the technical choices made.
- legal and standards expertise - complying with formal requirements may require involvement from legal experts, lawyers, or accredited experts
- deployment and social context expertise - an effective audit requires proper representation of outside interests and understanding of social harms of a system
- data creation and annotation - in most cases, benchmarks will need to be either created or modified for the evaluation
- computational resources - larger models and more complex systems require access to specialized hardware, and can incur substantive computational costs

These costs are be shared between the entities participating in the development of AI systems, the entities tasked with external audits, and broader participation from civil society. Developing entities need to follow internal accountability practices of documentation and testing. They typically have the technical expertise and computational resourced and engage in data creation in the course of developing an AI system.

Entities tasked with external audits have similar resource needs. We note however that where external audits can rely on work from a research community or other independent investigation into the audited systems, an efficient assesment can run with fewer costs.

We recommend paying particular attention to the cost of legal expertise to meet formal compliance requirements or obtain certification for standards — as those can be much more easily met by large companies than by academic or non-profit actors and Small and Medium Enterprises developing ML systems. Determining the right threshold for requiring a formal certification will be necessary for allowing different actors to participate in research and development of AI systems

Respectfully,
Yacine Jernite and Irene Solaiman
Hugging Face