HUGGING FACE, INC.
20 Jay Street, Suite 620
Brooklyn, NY 11201

HUGGING FACE

Hugging Face Information for NIST “Related to NIST’s Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence”

Authors: Led by Margaret Mitchell (Chief Ethics Scientist), with writing from Yacine Jernite (ML & Society Team Lead), Sasha Luccioni (Climate Lead), Clémentine Fourrier (engineering), Josef Fukano (security), Ezinwanne (Ezi) Ozoani (engineering), Irene Solaiman (Global Policy Lead)
Additional contributions from Imatag: Vivien Chappelier, Mathieu Desoubeaux

Submitted: Feb 2, 2024

Table of Contents

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. As part of our activity, Hugging Face provides social features and communication platforms for people interacting with AI systems, including social posts and discussion threads, in addition to hosting the AI systems themselves.

---

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

(1) Developing a companion resource to the AI Risk Management Framework (AI RMF)...

Information on "practices for implementing AI RMF core functions"

Addressing excerpt:

> - Risks and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness characteristics as defined in the AI RMF, as well as harms related to repression, interference with democratic processes and institutions, gender-based violence, and human rights abuses (see *https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/11/01/remarks-by-vice-president-harris-on-the-future-of-artificial-intelligence-london-united-kingdom*);
> - Current standards or industry norms or practices for implementing AI RMF core functions for generative AI (govern, map, measure, manage), or gaps in those standards, norms, or practices;
> - Recommended changes for AI actors to make to their current governance practices to manage the risks of generative AI;

We are thrilled by NIST's categorization of **govern, map, measure, and manage**. It aligns with some of the work we have prioritized at Hugging Face. Below we describe "practices for implementing AI RMF core functions", "gaps in those…practices", and "recommended changes…to manage the risks".

Govern

Data Governance

AI systems are a reflection of their data, and this data is often created by or about people. Data governance requires developers to account for the rights of these data subjects, including property, privacy, and user rights.

Governance practices that support respecting such rights depend on the specific development context and types of data used. For example, whether the AI system is developed as a commercial product or developed as a public good, or whether the data is obtained through web scraping or licensed through a commercial agreement with rights holders. These practices may involve gathering preference signals in whether specific items may be used for training (Spawning.ai), developing and applying new tools to remove privately identifying information (BigCode governance card), or providing tools to allow users of various online platforms to remove their data from training datasets (Am I In The Stack tool for GitHub opt-out).

In all cases, external stakeholders – including journalists and civil society organizations – have an important role to play in verifying that rights are indeed respected, provided they are given sufficient information about the workings or product of the data curation process. By outlining standards for what constitutes sufficient information to support external audits and data governance practices, NIST has an opportunity to foster responsible development of future AI systems.

In our work, we have outlined the requirements and proposed a data governance structure for web-scale text and media data. This work is primarily published in the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.

These ideas are based on a large workshop we co-led, focused on democratized LLM development, called "BigScience" (https://bigscience.huggingface.co), which included a Data Governance working group focused on how to best govern data across multiple governments and rights-holders.

Within the Data Governance working group, we identified the need to have a collaborative committee that works through different issues – the "Data Stewardship Organization" – with representatives for all stakeholders, including people who represent individuals' rights. See Figure 1 for a schematic of this structure.

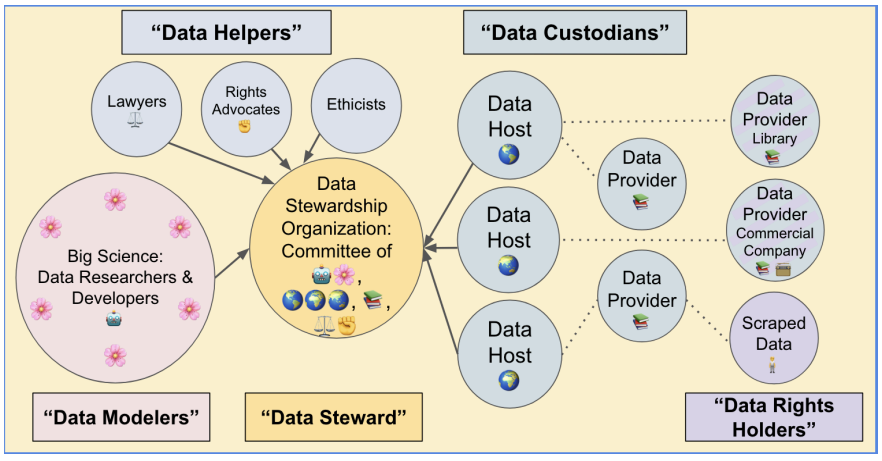

Figure 1. High-level schematic of Data Governance Structure
Source: https://dl.acm.org/doi/10.1145/3531146.3534637

Data Governance Roles

In addition to the central shared **Data Stewardship Organization**, there are the following roles:

- **Data Custodians**, who can be Data Hosts, Data Providers, or both, using the data from Data Rights Holders.
    - **Data Hosts** make the data available for analysis
    - **Data Providers** are individuals or institutions who have text, image, or audio data that they can legally make available.
    - **Data Rights Holders** are the creators or owners of the data
- **Data Modelers** are researchers and developers who use the data.

Agreements and contracts are necessary between these different entities to best represent the different stakeholders in AI data usage.

Platform Governance

We ground governance of the Hugging Face platform on key values. These include **openness**, **transparency,** and **collaboration**. We also strive for **kindness**. Decisions regarding everything from design to reporting processes use these values as a foundation.

In order to govern such an open community platform and maintain key values, it is necessary to have a Code of Conduct (see: https://huggingface.co/code-of-conduct), which also describes different actions that the organization may take in different situations, and draws from what is appropriate with respect to Content Guidelines (see: https://huggingface.co/content-guidelines). A cornerstone of our Code of Conduct and Content Moderation policies is also **consent**, a key value for inclusion that also addresses many kinds of problematic content that may be shared.

Map

For training datasets of ML systems, organizations should perform the following mapping functions:

- **Mapping stakeholder groups:** organizations should work on identifying all relevant groups of data subjects (people who create or are represented in the dataset) as well as algorithm subjects (people whose lives will be affected by models leveraging the dataset through the training and deployment of the full AI systems).
- **Mapping stakeholder rights:** organizations should know what regulations might apply to the use of data and algorithmic decision systems. We draw particular attention to relevant sectoral regulation in health, education, and finance, which governs how information may be used and where data subjects are entitled to notification or explanation regarding their data use. Organizations should make sure to look to federal regulations and state regulations on those topics, including relevant rules on privacy or customer protection.

- **Mapping flows of information:** finally, organizations should provide a mapping of the flow of information that supports their activity, for example in the form of a data management plan. In current Machine Learning development practice, data gathered for an initial purpose is often re-used outside of its initial context, which may have consequences for the identification of stakeholder groups and rights outlined above.

We provide an example of mapping between stakeholders and rights in Table 1 below. This was created by centering on the pillars of development defined in Manage (Figure 2) and articulating the groups involved in each.

Example Stakeholder Groups	**Relevant Pillars** (from Manage & Figure 2)	**Example Rights Affected** (adapted from the UDHR)
Data creators and data subjects, including those producing "raw" data (such as artists), those annotating it (such as crowdworkers), and the people represented in the data	Data Collection Training Processes	- Right to Gain a Living - Right to [Intellectual] Property
AI developers, which may be individual engineers or larger organizations (such as tech companies)	Data Collection Training Processes Model Evaluation & Analysis System Deployment	- Right to Gain a Living - Right to [Intellectual] Property
AI deployers, who leverage the technology for different applications (such as companies and government agencies)	Model Evaluation & Analysis System Deployment	- Right to Gain a Living - Right to [Intellectual] Property
AI users, who interact with the technology made available by deployers (such as people in education, healthcare, and finance)	System Deployment	- Freedom from Harm - Freedom of Expression - Right to Privacy - Right to Non-Discrimination
AI-affected, who may have AI technology applied to them, whether or not they chose to (such as in surveillance)	System Deployment	- Freedom from Harm - Right to Privacy - Right to Non-Discrimination - Rights of the Child

Table 1. Example stakeholder groups corresponding to each pillar include the following, which is non-exhaustive and not mutually exclusive.
Also included are a set of example rights for each; see the linked UN article for further detail.

For each pillar, we can derive the benefits, harms, and risks to different people in different AI contexts, identifying the potential for positive and negative impact, and which rights may be affected.

Measure

Data Measurement

This concept is defined in Mitchell et al. (2022) Measuring Data.

This is the key idea: Just as we *evaluate* models, we should *measure* data.

Prioritizing/working on data measurement with the same rigor and frequency as model evaluation would be a significant change that we recommend AI actors must make in order to better identify foreseeable outcomes – risks and benefits – from systems trained on that data.

For example, measuring data allows us to identify, and quantify the risk of, things that a model might learn and emulate. This includes measuring things such as:

- Stereotypes
- Sentiment, opinions and stances
- Persuasion tactics
- Incorrect information, which could be realized by a model as misinformation, or abused to create disinformation
- Hate and toxicity
- Private and personally identifiable information, using raw counts of different types

These are measurable using:

- Statistical techniques that do not require additional annotations, such as co-occurrence statistics (e.g., correlation and association metrics) or frequency statistics (e.g., tf-idf, fit to Zipf's law)
    - We have implemented a proof-of-concept at https://huggingface.co/blog/data-measurements-tool.
- Real-valued scores from classifiers, such as sentiment or toxicity classifiers that provide both the polarity of the content and its predicted magnitude;
- Human annotations
    - Such as described in ISO work on "data quality measures" (e.g., ISO 25024).

Data measurement can also be applied to the **output of models** – think of model output as its own dataset – as a way to measure different model properties. For example, one can measure learned stereotypes from a generative model without additional annotations (see https://dl.acm.org/doi/10.1145/3461702.3462557) by generating a large amount of content from the model and measuring the co-occurrence between different items in its generations, like *woman* and *smile*. We can combine this with human input on social categories and derive further insights as well (see https://aclanthology.org/2023.acl-long.84.pdf).

Further, the company **Lilac ML** (https://www.lilacml.com) has prioritized exactly this kind of work, and may be a great partner for NIST to help provide tools for responsible data analysis.

Model Evaluation

The rapid proliferation of models, architectures and modalities makes it important to have reliable ways for comparing and evaluating models. Depending on the context of development, different metrics and benchmarks should be used and reported in accompanying reports and literature. Given the open-endedness of generative AI models, there is often no single correct answer, making evaluation even more difficult and resulting in a plethora of different ways to evaluate new and existing models. Hugging Face's evaluate Python package and Eleuther AI's language model evaluation harness are examples of tools that aim to make model evaluation more standardized and systematic.

Also see our section on Benchmarking Results.

Manage

An approach that helps to manage the risks of AI (including generative AI) within a governance framework requires modularizing the end-to-end development of AI systems and engaging in critical analyses for each. A core part of AI systems are machine learning models, which can be modularized into 4 components (see Figure 2). Each component requires robust analysis articulating risks, rights, and proactive solutions.

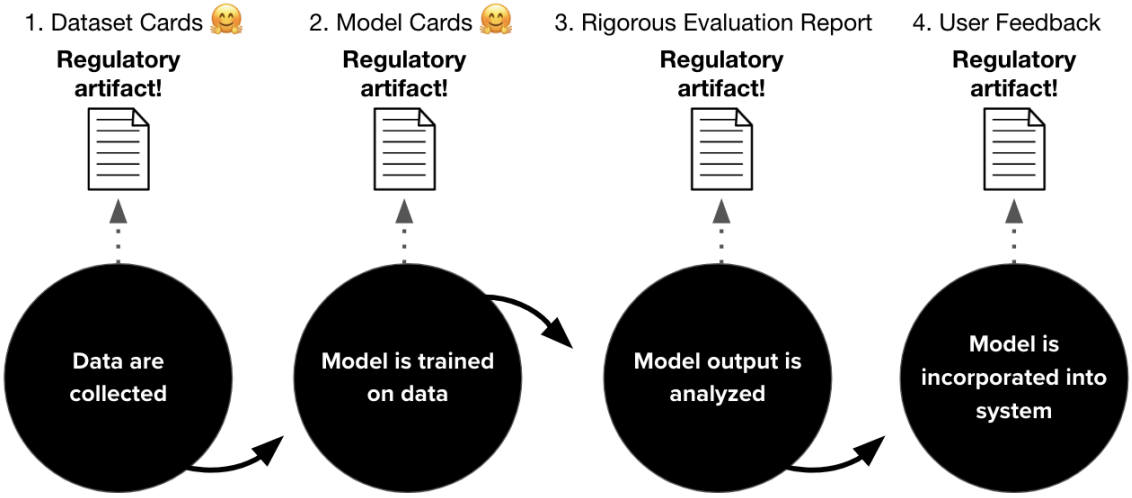

Figure 2. Regulatory artifacts in an end-to-end ML development pipeline

We discuss the regulatory artifacts of (1) Dataset Cards and (2) Model Cards in Information on "forms of transparency and documentation" below. Here, we detail (3) Rigorous Evaluation Report and (4) User Feedback.

Rigorous Evaluation Report

The Rigorous Evaluation Report requires **disaggregated evaluation**. Evaluation is said to be "disaggregated" when it is not one single, "aggregate" score, but a set of scores for different slices, such as scores for different subpopulations in different contexts. A model is said to be "fair" with respect to a characteristic (e.g., "fair across genders") when the evaluation metrics are equal across the different subpopulations for that characteristic (e.g., when the model's evaluation score for men, the model's evaluation score for women, the model's evaluation score for nonbinary, etc., are all equal).

Our experience suggests that Rigorous Evaluation Reports are created well by first following these steps:

> **Step 1.** Define the relevant people (stakeholder groups and subpopulations).
>
> **Step 2.** Identify how each group may be affected in different contexts.
>
> **Step 3.** Determine the metrics and measurements to evaluate and track the effects on each group (Step 1) in each context (Step 2).
>
> **Step 4.** Evaluate model performance with respect to the groups and contexts (Step 3).

For Steps 1 and 2, the approach requires crossing people by contexts. "People" are split into *users* and *those affected*, *intended* and *unintended*. "Contexts" are similarly split into *intended* and *unintended*, as well as *out of scope*. This can be seen as primarily a 2x4 grid, where each cell must be filled out – see Figure 3.

		People			
		Users		**Those affected**	
		Intended	**Unintended** Both malicious actors & people un-accounted for in development	**Intended**	**Unintended** Both people in training data & people the technology is used on
Use Contexts	**Intended**	Beneficial technology		Beneficial technology	
	Unintended Both harmful contexts & those unmodeled in development		Problematic technology		Problematic technology
	Out of scope	Technology won't work			

Figure 3. Foresight in AI chart. When developing responsibly, it should be possible to fill out each of these cells.

This means that rigorous evaluation requires first answering questions the cells correspond to

such as *what are the use contexts, and who is involved in these contexts? What are the intended or beneficial uses of the technology in these contexts? What are the unintended or negative ones?*

Clearly defined subpopulations and use contexts can inform the selection of metrics to evaluate the system in light of foreseeable outcomes.

User Feedback

Those affected should be able to easily provide feedback: In order to know how well the system is working, and to have serious errors immediately flagged, there must be a simple user-facing mechanism for immediate feedback. At Hugging Face, we have implemented this in several ways:

1. A "Report this" button that appears alongside data, models, and demos – and the report can either be public or anonymous
2. A "Community tab" for open discussion alongside different artifacts.

Information on "roles"

Addressing excerpt:

> - The types of professions, skills, and disciplinary expertise organizations need to effectively govern generative AI, and what roles individuals bringing such knowledge could serve;
> - Roles that can or should be played by different AI actors for managing risks and harms of generative AI (*e.g.,* the role of AI developers vs. deployers vs. end users);

ML Lifecycle Roles

Please see our section on the AI RMF Map pillar, Table 1.

Data Development Roles

Details on the roles, professions, skills, etc., needed for data governance were described in the above sections addressing the AI RMF Govern pillar (esp. Data Governance and Data Governance roles).

There are also different types of responsibilities needed for **dataset development** (including creation and maintenance). These roles should serve to produce the artifacts listed in Figure 4, which are needed for responsible data development.

Artifact	Details
Dataset Requirements Spec	Target Properties Intended uses
Dataset Design Doc	How dataset will be collected
Dataset Implementation Diary	Status of collection attempts How issues were solved
Dataset Testing Report	Measurement of dataset Results of adversarial examination Issues in dataset
Dataset Maintenance Plan	Updates for opt-out Handling stale data

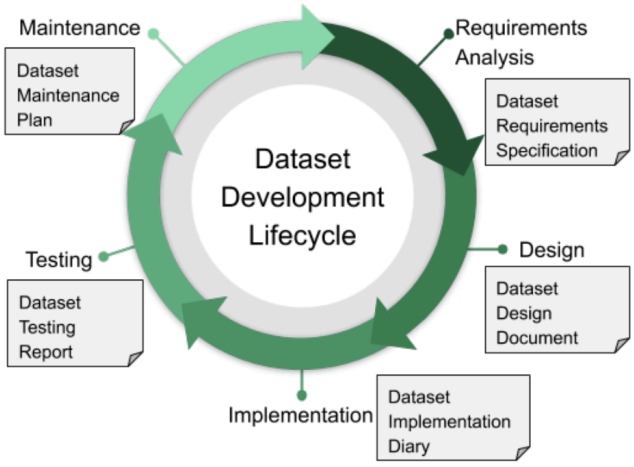

Figure 4. Roles and Responsibilities for Dataset Development
Source: https://dl.acm.org/doi/pdf/10.1145/3442188.3445918

Please also see the suggestions in *Khan and Hanna. 2022. "The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability."*

Model Roles

Roles in responsible governance of AI models should include the appropriate diversity needed to identify relevant information about a machine learning model for different kinds of audiences.

This includes the following (adapted from our Annotated Model Card, https://huggingface.co/docs/hub/model-card-annotated). One person may have more than one role:

- The **developer**, who writes the code and runs training;
- The **sociotechnic**, who is skilled at analyzing the interaction of technology and society long-term (this includes lawyers, ethicists, sociologists, or rights advocates);
- The **project organizer**, who understands the overall scope and reach of the model, and who serves as a contact person to connect different parties.

The **developer** is necessary for providing information about a model's training procedure and technical specifications. They are also useful for identifying technical "limitations", such as likely failure modes, and additionally may be able to provide recommendations for addressing those limitations. They are responsible for calculating and providing results of model evaluation, working with the other roles to define the most appropriate rigorous evaluation.

The **sociotechnic** is necessary for identifying different contexts where a model may be used, the different subpopulations that may be differently affected by the model, and how the model may be used in different contexts or with respect to different subpopulations. The types of bias a model might have, the risks different applications bring, and out-of-scope model usage naturally fall out of this kind of analysis.

The **project organizer** is necessary for identifying content such as basic uses of the model, licensing requirements, terminological alignment, etc.

Information on "current techniques and implementations"

Addressing excerpt:

> - Current techniques and implementations, including their feasibility, validity, fitness for purpose, and scalability, for:
>   - Model validation and verification, including AI red-teaming;
>   - Human rights impact assessments, ethical assessments, and other tools for identifying impacts of generative AI systems and mitigations for negative impacts;
>   - Content authentication, provenance tracking, and synthetic content labeling and detection, as described in Section 2a below; and
>   - Measurable and repeatable mechanisms to assess or verify the effectiveness of such techniques and implementations.

Identifying impacts and developing mitigations

Channels for communication should be maintained for responsible disclosure of security issues, vulnerabilities, bias, and inappropriate content. For fully public-facing models, the mechanisms to provide feedback must be open for all of the public (even if what is reported is not made visible due to security concerns). For more private or internal systems, channels must be open for private/internal employee communication.

At Hugging Face, we have implemented such mechanisms, described in our section on operationalising User Feedback as part of NIST's Manage pillar.

Associated with these channels should be a procedure to document, triage (escalate), risk rank, and resolve items reported to the organization using AI. (Note that mature organizations will generally already have a similar pattern they can follow from bug bounty, security operations, privacy data subject requests, and related programs.) **"Escalations"** or triaging occurs when there is an immediately pressing issue that must be addressed. In these cases, people throughout the company are brought in to quickly put their heads together to define the **immediate**, **short-term**, **medium-term**, and **long-term** solutions, along with who is the "point person" for each.

Additionally, companies that rely on or invest in these models for business activities should consider **bounty initiatives** for the responsible disclosure of model issues.

Ongoing automated and periodic manual testing of models is also required to ensure they operate within expected parameters. When using AI models, **subject matter experts** should be employed or consulted to use and review results on an ongoing basis. For example, if a company builds a fraud detection system, they should still have a fraud specialist that can interpret results and identify anomalies. This is an 'external to the system' control that provides the opportunity to identify problems with the output and then investigate where the problems occurred upstream. These subject matter experts will complement automated scans of data and outputs to ensure they fall within predefined thresholds.

Assessments

Please see our section on Auditing.

Content authentication, provenance tracking, and synthetic content labeling and detection

We believe that content provenance for AI datasets, models, and systems, is critical for the future of responsible and ethical AI.

Models and systems

For models and systems, we are entering a world where it's becoming unclear which content is created by AI systems, and impossible to know where different AI-generated content came from. Bad-faith actors can further compound the issue, using this new technology to deceive, misrepresent, and influence people without a trace. These issues are directly addressed by **embedding content provenance information in generated content, using techniques such as watermarking**, which helps us to know what has been created by an AI system and where it came from. It provides for mechanisms that help the public gain control over the role of generated content in our society.

We have collaborated with multiple parties to demonstrate the state of the art in content provenance mechanisms. This includes:

- **Truepic** (https://hf.co/blog/alicia-truepic/identify-ai-generated-content), a leading provider of authenticity infrastructure for the internet who has demonstrated how to:
    a. Cryptographically secure metadata into any generated image using the open C2PA standard: https://huggingface.co/spaces/Truepic/ai-content-credentials
    b. Generate "invisible QR codes" as image watermarks that can be used to retrieve further image metadata: https://huggingface.co/spaces/Truepic/watermarked-content-credentials
- **Imatag** (https://hf.co/blog/imatag-vch/stable-signature-bzh), who specializes in digital watermarking and has demonstrated how to embed secure and robust invisible watermarks during the image generation process: https://huggingface.co/spaces/imatag/stable-signature-bzh

Verifying the connection between data and models

How to **verify** what data a model has been trained on is currently open research (see https://openreview.net/forum?id=TwLHB8sKme). While some developers may report the data they used, there are limited ways to prove whether this is true, or exhaustive.

Measurable and repeatable mechanisms to assess or verify the effectiveness of such techniques and implementations.

Some approaches and resources that can serve as starting points for delving deeper into this include:

- **Bounty Catching**: The cybersecurity field has illustrated success in Bug Bounty Programs. Some notable examples include:
    - HackerOne: HackerOne is one of the leading platforms for running bug bounty programs. They release detailed blog posts, where they provide insights into measuring the performance of their internal and business collaborative bug bounty initiatives using key indicators like the total number of reports received, average time to resolution, and severity distribution of reported bugs. By focusing on these factors, organizations can gauge the efficiency of their escalation testing processes and continuously improve them.
        - https://www.hackerone.com/vulnerability-and-security-testing-blog
    - Synack: Synack is a crowdsourced security platform that shares several quantifiable measures to determine the efficacy of a bug bounty program. Some of these metrics include the number of unique vulnerabilities discovered, time-to-fix, and cost savings compared to traditional penetration testing methods. Regularly monitoring these parameters enables continuous improvement and fine-tuning of escalation testing models
        - https://www.synack.com/wp-content/uploads/2022/09/Crowdsourced-Security-Landscape-Government.pdf

- **Subject Matter Experts:** Organizations and governments in the U.S. are continuing to advance AI initiatives by creating initiatives on collaboration with subject matter experts. By learning from prior efforts, these initiatives succeed in delivering accurate predictions, innovative tools, and ethical standards. This includes:
    - **The American Heart Association**: Recently announced the launch of its precision medicine platform, which utilizes AI and machine learning to analyze genomic, biological, and lifestyle data. Medical experts contributed to the design and validation of the platform, guaranteeing medical relevancy and ethical standards.
        - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8452247/
        - https://www.ahajournals.org/doi/full/10.1161/CIRCOUTCOMES.121.007949
    - **NASA's Frontier Development Lab**: Hosted summer fellowships for space scientists and AI researchers to collaboratively tackle challenging scientific questions. Participants exchange knowledge, creating AI tools and applications that push the frontiers of both fields.
        - https://frontierdevelopmentlab.org/

Furthermore, combining insights from separate disciplines can often lead to innovative solutions for novel problems. Cross-referencing these examples may inspire new ideas and approaches for evaluating AI techniques and implementations.

Information on “forms of transparency and documentation”

Addressing excerpt:

> - Forms of transparency and documentation (*e.g.,* model cards, data cards, system cards, benchmarking results, impact assessments, or other kinds of transparency reports) that are more or less helpful for various risk management purposes (*e.g.,* assessment, evaluation, monitoring, and provision of redress and contestation mechanisms) and for various AI actors (developers, deployers, end users, etc.) in the context of generative AI models, and best practices to ensure such information is shared as needed along the generative AI lifecycle and supply chain);

At Hugging Face, we have worked extensively on operationalizing “model cards” and “dataset cards” for all models and datasets that are shared on the platform.

Model Documentation

We have spent years working with organizations and people throughout the world to fill out model cards. This effort has been led in part by the lead author on the Model Card paper, Margaret Mitchell. Her writing on **what people are missing** when creating Model Cards is available here: https://www.techpolicy.press/the-pillars-of-a-rightsbased-approach-to-ai-development/, in “Deep Dive: Model Cards” section. Critically, people are generally missing **disaggregated evaluation** in their reporting of results, which requires first identifying the relevant subpopulations to evaluate

model performance on; then identifying the types of errors likely to cause harm; then defining evaluation metrics based on these errors; and finally applying the evaluation with the selected metrics across the disaggregated subpopulations.

However, since skipping disaggregated evaluation is becoming increasingly common, it may make sense for Model Cards to exclude this information and have it instead reported in a Rigorous Evaluation Report, as described in our section on the AI RMF Manage pillar..

For further detail, here is our guidebook on Model Cards: https://huggingface.co/docs/hub/main/model-card-guidebook.

This covers:

- How to fill the model card out, including roles and responsibilities (https://huggingface.co/docs/hub/main/model-card-annotated)
- Model card user studies (https://huggingface.co/docs/hub/main/model-cards-user-studies)
- A landscape analysis of ML documentation tools (https://huggingface.co/docs/hub/main/model-card-landscape-analysis)

Dataset Documentation

We have worked on developing new standards and tools to support dataset transparency, and have written extensively on the current state of possibilities and practices in the field (https://huggingface.co/blog/yjernite/data-transparency) Our work draws from approaches such as “Datasheets for Datasets” (https://arxiv.org/abs/1803.09010).

Understanding the make-up and characteristics of the datasets used in AI systems is **essential to fully controlling the development of AI**: AI model behavior is based on the data it is trained on. Dataset documentation should provide information surrounding data provenance as well as high-level statistics about datasets, such as the languages they contain and the most common words or categories. Providing this information in an easily accessible (and machine-readable) format can help people to understand which datasets are useful for which tasks. This can also help improve the representativity of datasets, for instance to highlight languages that are under-represented. Detailed documentation may also take the form of **metadata**, with information about each instance in the dataset. This can (among other things) support opting out of datasets for data creators or rights-holders who do not want their data to be distributed or used for training AI models.

Dataset documentation sits within a set of data mechanisms that ensure that the risks of the technology are properly managed - including risks of discrimination, risks to privacy, and broadly making sure that the technology follows all applicable regulations. This set of mechanisms include:

- Direct access to the datasets by some categories of stakeholders

    - See our sections on the AI RMF Map pillar, Data Development Roles, Data Governance Roles; and our sections on Auditing for information on stakeholders external to the data development process.
- Sufficient public-facing documentation according to broadly applicable standards, as described here
- Interactive exploration tools to support investigation from actors outside of the development chain
    - One recent example is Lilac ML's "Garden" (https://docs.lilacml.com/blog/introducing-garden.html).
    - Further tools are linked in our blog post on data transparency (https://huggingface.co/blog/yjernite/data-transparency).

For example, we note that:

- The propensity of a Large Language Model to produce hateful or toxic text is directly correlated with the presence of this kind of content in the training dataset. At the same time, efforts to automatically remove this kind of text have been shown to introduce their own biases, and have a disparate impact on marginalized groups (https://huggingface.co/papers/2104.08758). While direct access to a training dataset is not always possible, listing the original sources of the dataset and providing documentation of the specific mitigation techniques leveraged to reduce the proportion of hateful or toxic text in the final dataset, including the specific criteria and thresholds used in the filtering, can help ensure that risks tied to hate speech and risks of discrimination are properly balanced (further described in the recent preprint at https://arxiv.org/abs/2401.06408).
- Similar considerations also apply to other content modalities, including for image generation systems, where efforts to limit risks of generating violent or other kinds of undesirable content (e.g., https://openai.com/research/dall-e-2-pre-training-mitigations) have also led to decreased social diversity in the model outputs (as we demonstrate in our NeurIPS paper on "Stable Bias", https://arxiv.org/abs/2303.11408).
- Additionally, over-estimating a model's performance on specific tasks is an important risk factor of AI deployment - especially when the model is in a position to significantly shape people's access to services or be integrated into infrastructure. One of the best studied sources of over-estimation is benchmark contamination (described in https://arxiv.org/abs/2312.16337), where a model may be evaluated on a benchmark that it was fully or partially trained on.
    - While decontamination approaches can help developers, downstream adopters of the technology may not have sufficient information about the training data to safely perform their own evaluation and are at risk of spending significant effort on testing a model's safety in a controlled setting only to have it fail in real-world deployment because they could not meaningful separate their safety benchmarks from the model's training dataset.

As a result, we recommend:

- For small to medium datasets, for example datasets from a single or from a few identified sources of up to tens to hundreds of data items (sentences, images, documents, etc.), organizations should at the very least provide documentation of the dataset in the form of a data statement, datasheet, or data nutrition label.
- For larger datasets, including composite datasets and web-crawled datasets, organizations should further provide documentation of the major sources used for training models, and individual documents covering all major sources following one of the standards mentioned above.
- Further, any automatic risk mitigation strategies at the dataset level should be sufficiently documented to allow external stakeholders to evaluate the trade-offs and values it prioritizes, including choices motivating data filtering.

Assessment and Evaluation

Benchmarking

We have found leaderboards to be extremely effective and influential in benchmarking and reproducing results. In particular our Open LLM Leaderboard (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard) garners 250K visits monthly, with more than 4K models submitted and evaluated, and shows how well different open LLMs perform across a variety of well-known benchmarks. Our teams also produced an LLM performance leaderboard (https://huggingface.co/spaces/optimum/llm-perf-leaderboard) to evaluate energy efficiency and consumption of LLMs at inference time, and the Big Code Leaderboard (https://huggingface.co/spaces/bigcode/bigcode-models-leaderboard), to evaluate model's programming quality.

We have provided support for a number of more specialized leaderboards, in collaboration with universities or companies:

- The Chatbot Arena Leaderboard, which uses human ratings to compute an Elo score of available (open and closed) models (https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard)
- The LLM Safety Leaderboard, looking at toxicity, bias and robustness, among others (https://huggingface.co/spaces/AI-Secure/llm-trustworthy-leaderboard)
- Two Hallucinations Leaderboards, to evaluate the propensity of models to say incorrect things (https://huggingface.co/spaces/hallucinations-leaderboard/leaderboard and https://huggingface.co/spaces/vectara/Hallucination-evaluation-leaderboard)
- Several specialized leaderboards for specific languages, like Korean, Dutch, etc …

Social Impact

To determine the impact of AI systems on people, especially in safety and capabilities that can cause harm, evaluations targeted at social impact can give insight on certain model behaviors. Social impacts of systems can be evaluated by two means: technical evaluations of the model and by surveying users and affected populations.

Technical evaluations for the social impact of models include quantifying harmful biases, toxicity, disparate performance of language, and environmental cost of training and inference. However, some social impacts, such as overreliance on outputs and trust in information, can only be measured by surveying and/or interviewing affected people. Technical evaluations of social harms (such as biases) can also be overly narrow, with reductive identity group categorization such as solely "profession".

That said, social impact assessment is an area of open research. Methods and tools needed to evaluate complex social harms have many limitations, from human evaluator bias, to classifier biases, to being outdated for social norms. The timelines needed for examining social aspects of safety can also vary based on using time as a variable to analyze, e.g. in trust in information over time or impact on human labor and the global economy.

Information on "watermarking"

Addressing excerpt:

> - Economic and security implications of watermarking, provenance tracking, and other content authentication tools;
> - Efficacy, validity, and long-term stability of watermarking techniques and content authentication tools for provenance of materials, including in derivative work;

Also see our response on Content authentication, provenance tracking, and synthetic content labeling and detection.

Writing in this section provided in part from Imatag (https://www.imatag.com), who specializes in digital watermarking. Their solutions are some of the only independent watermarking technologies for AI models. Their full response on this section is attached in the Appendix.

Watermarking is a technique designed to unobtrusively mark content in order to convey additional information, such as authenticity. Within AI, watermarking involves adding machine recognizable patterns to digital content (such as images), and these patterns convey information such as where the content came from and whether it's synthetic. New AI/digital watermarking methods are also designed to be **imperceptible** to people, so as not to disrupt the content but still make it possible to detect and track digital content as soon as it's shared. New digital watermarking methods are increasingly **robust** to common alterations (e.g., compression, cropping, and color changes in the case of images), and AI watermarking solutions are said to be "**secure**" when malicious attempts to remove, extract, or forge watermarks are not possible

from a third-party unaware of the secret (i.e., the "key" or "cipher") used to protect the watermarking solution.

Two distinct methods are currently being studied for watermarking AI-generated images. The first method involves watermarking the output of the generative AI model **after it's created**, just as can be done with any content we might upload online. In this context, there is evidence that it's possible to create digital watermarking that is fast, robust and secure, for closed systems: The company Imatag is delivering such a system for newswire companies (https://www.prnewswire.com/news-releases/afp-selects-imatag-to-track-the-uses-of-its-photographs-301550539.html). However, applying watermarks as a post-process in an open system has the strong drawback that it is easily removed by simply commenting out some of the code.

The second method implements the watermarking process **during the creation** of AI content. This is **possible to securely apply to open AI models**, such as those made available on the Hugging Face platform. This enables the distribution of AI models that are already modified to automatically embed watermarks as part of their generation process. This lowers the burden on individual developers to add on their own watermarking solutions, and provides secure watermarking by "default".

The trade-offs between imperceptibility, robustness, and security is key to evaluating AI/digital watermarking systems. As digital content spreads on the Internet, it is often modified multiple times for technical or editorial reasons. For example, it may be saved from a screenshot, saved as different file types, recompressed, or cut off. This is why current open-source watermarking solutions are not robust enough to these kinds of alterations for practical use (see https://medium.com/@steinsfu/stable-diffusion-the-invisible-watermark-in-generated-images-2d68e2ab1241, https://www.wired.com/story/artificial-intelligence-watermarking-issues/). However, it is important to recognize that although watermarking is not perfect, we should not let perfect be the enemy of the good: The use of watermarking will help good actors and mitigate many bad actors. (Article where we further discuss this available here: https://venturebeat.com/ai/invisible-ai-watermarks-wont-stop-bad-actors-but-they-are-a-really-big-deal-for-good-ones/)

Information on "disclosing errors"

Addressing excerpt:

> - Criteria for defining an error, incident, or negative impact;
> - Governance policies and technical requirements for tracing and disclosing errors, incidents, or negative impacts;

Within code development, errors can be easily tracked and traced using **git protocols and tools** (see https://git-scm.com) or similar versioning software that can:

- Track who did what, when
- Flag issues before merging/incorporating new code into a shared code repository.

Within a broader community, it's also possible to create mechanisms for community feedback. At Hugging Face we have adopted an "open" approach such that this feedback is viewable and accessible to everyone side-by-side with models and data (called the "Community Tab"; see https://huggingface.co/blog/community-update), so those developing and using these assets can also interact with others. Because community feedback is open for everyone for each model, dataset, or demo, **people affected** – not just direct users – can provide information about the effect of the systems. This mechanism is also mentioned in our response above on User Feedback as part of NIST's Manage pillar.

In order to govern such an open community feedback system, see Platform Governance above.

(2) Creating guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities and limitations through which AI could be used to cause harm.

Information on "auditing AI"

Companies that use AI for business purposes should consider performing internal and external audits that correspond to the risks that AI poses to the business as well as consumers, data subjects, and industry. Such risks may be informed by the **sociotechnic** role described in our above response on roles. Internal audits should use a combination of manual tests and automated / algorithm based analysis. Audits can also be **"second-party"** – open to a single person or an organization under NDA – or **"third party"**, open to people without connections to the company. See "Missing links in AI governance (UNESCO)" for more information on these kinds of audits.

These audits should examine the **regulatory artifacts** described in our section on the AI RMF Manage pillar, as well as the processes to produce them, in order to verify the work appropriately addresses different values, processes, or laws. The regulatory artifacts are aligned with standard tech development practices and so provide for a clear "connection point" between tech companies and tech auditors.

Information on "AI Evaluations"

Addressing excerpt:

> - Definition, types, and design of test environments, scenarios, and tools for evaluating the capabilities, limitations, and safety of AI technologies;
> - Availability of, gap analysis of, and proposals for metrics, benchmarks, protocols, and methods for measuring AI systems' functionality, capabilities, limitations, safety, security, privacy, effectiveness, suitability, equity, and trustworthiness.
>
> ...

- Generalizability of standards and methods of evaluating AI over time, across sectors, and across use cases;
- Applicability of testing paradigms for AI system functionality, effectiveness, safety, and trustworthiness including security, and transparency, including paradigms for comparing AI systems against each other, baseline system performance, and existing practice

Please see our section on Assessment and Evaluation.

Information on "AI Red-Teaming"

Addressing excerpt:

- Use cases where AI red-teaming would be most beneficial for AI risk assessment and management;
- Capabilities, limitations, risks, and harms that AI red-teaming can help identify considering possible dependencies such as degree of access to AI systems and relevant data;
- Current red-teaming best practices for AI safety, including identifying threat models and associated limitations or harmful or dangerous capabilities;
- Internal and external review across the different stages of AI life cycle that are needed for effective AI red-teaming;
- Limitations of red-teaming and additional practices that can fill identified gaps;
- Sequence of actions for AI red-teaming exercises and accompanying necessary documentation practices;
- Information sharing best practices for generative AI, including for how to share with external parties for the purpose of AI red-teaming while protecting intellectual property, privacy, and security of an AI system;
- How AI red-teaming can complement other risk identification and evaluation techniques for AI models;
- How to design AI red-teaming exercises for different types of model risks, including specific security risks (e.g., CBRN risks, etc.) and risks to individuals and society (e.g., discriminatory output, hallucinations, etc.);
- Guidance on the optimal composition of AI red teams including different backgrounds and varying levels of skill and expertise;
- Economic feasibility of conducting AI red-teaming exercises for small and large organizations; and
- The appropriate unit of analysis for red teaming (models, systems, deployments, etc.)

"Red-teaming" applies after a model has already been trained, and as such is a post-hoc approach to AI safety. It operates similarly to "whackamole": Given an already problematic system, try changing one component and seeing what other issues may pop up. It is a relatively accessible means of testing a system for subject matter experts who can query a system through low barrier interfaces. More technical details on red-teaming at Hugging Face is available at https://huggingface.co/blog/red-teaming.

It is also one of the weaker approaches for guaranteeing safety, as it does not holistically assess an AI system's safety and cannot influence model behavior in the way that other methods do, such as careful data curation. As one of many tools in the Responsible AI toolbox, it is one of the last that can be applied to identify model harms within the chain of model development (Figure 2), preceding user feedback and external audits.

Summarizing from above, the "red-team" approach thus applies at the tail end of the following roughly consecutive interventions:

1. **Public Consultation:** Before beginning to train a model, best practices include consulting with experts who are mostly likely to use or be affected by the model in order to understand what to prioritize in data collection and model training. This also serves as input for impact assessments.
2. **Data requirements:** Specify what is desired from the model and collect data in light of these goals, as described in Information on Roles, Data section.
3. **Data analysis and measurement:** Identifying issues of representation, stereotype, etc., as described in Information on NIST's Measure pillar, Data Measurement section.
4. **Model training, mapping inputs to outputs:** Analyzing the effect of different training data slices on different model behaviors, and excluding those data instances that result in problematic behavior.
5. **Model training, disaggregated evaluation:** As the model trains, different model checkpoints can be evaluated with respect to different areas of concern.
6. **Model convergence, disaggregated evaluation:** When the model is done training, disaggregated evaluation can similarly be used to identify harms with respect to different contexts and subpopulations.
7. **Red-Teaming:** We recommend https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/.
8. **User feedback:** Including bug bounties, as described in our section on Model Validation & Verification, including red-teaming.

2. Reducing the Risk of Synthetic Content

Information on “synthetic content”

Addressing excerpt:

> Existing tools and the potential development of future tools, measurement methods, best practices, active standards work, exploratory approaches, challenges and gaps are of interest for the following non-exhaustive list of possible topics and use cases of particular interest.
>
> - Authenticating content and tracking its provenance;
> - Techniques for labeling synthetic content, such as using watermarking;
> - Detecting synthetic content;
> - Resilience of techniques for labeling synthetic content to content manipulation;

Please see our responses on Watermarking and Content Authentication.

Addressing excerpt:

> - Approaches that are applicable across different parts of the AI development and deployment lifecycle (including training data curation and filtering, training processes, fine-tuning incorporating both automated means and human feedback, and model release), at different levels of the AI system (including the model, API, and application level), and in different modes of model deployment (online services, within applications, open-source models, etc.);
> - Testing software used for the above purposes; and
> - Auditing and maintaining tools for analyzing synthetic content labeling and authentication.

Synthetic content is problematic in part because it can be used for disinformation and non-consensual sexualization. For both, the **platform** where the content would be distributed has a critical role to play (a point also addressed in our section on Organizational Solutions for non-consensual intimate imagery). Platforms should scan shared content to verify whether or not it is synthetic, and alert users accordingly. One example of how to do this would be:

- Organizations that make generative AI models available embed metadata and/or invisible watermarks within the generated content.
- For watermarks, this might be using a proprietary method that would have corresponding proprietary detection software.
    - Major platforms where synthetic content is shared can then run the different available proprietary detection software tools on shared content to identify whether any of them have additional metadata that can be used to change how the content is shared on the platform.

Information on “non-consensual intimate imagery”

Addressing excerpt:

> - Preventing generative AI from producing child sexual abuse material or producing non-consensual intimate imagery of real individuals (to include intimate digital depictions of the body or body parts of an identifiable individual);

There are multiple points of intervention for combating CSAM and non-consensual intimate imagery. These can be roughly categorized as “technical solutions” and “organizational solutions”

Technical Solutions

Text-to-image systems

CSAM images and non-consensual sexualization can be generated as a response to **prompts**, meaning texts that a user types in. These queries can either be **seeking prompts** or **non-seeking prompts**. These terms can also apply to broader non-consensual content, including sexual and violent material.

Seeking prompts can be identified utilizing a text classifier to determine whether sexualized words are being used, and whether they are co-occurring with terminology for children or other people. Critically, these tools must be robust enough to handle misspellings, extraneous characters, etc.

Non-seeking prompts are those that return sexualized imagery even though the user has not requested them. Potential inappropriate sexual image content can be identified by running classifiers on the returned images, and not showing the image is problematic content is detected.

Multimodal (text prompt and image) classification is also an option, training a system to jointly recognize whether the prompt, the image, or both are sexualized.

Image-only systems

Some generative image approaches make it terrifyingly easy to generate CSAM. For example, new techniques in generative AI [information provided privately] can create new types of personalized images more easily. Identifying problematic content involves running detection algorithms on generative images, either during generation or once the image is already generated.

Organizational Solutions

One place where problematic content proliferates is on platforms, such as social media platforms. **Platforms have a critical role to play in combating proliferation**. A shared image can

embed an invisible watermark using common software, which the platform can then also use for detection. Also see our information on watermarking and information on synthetic content for further details how this can work.

Accountability can also be directed to the person conducting the generation and the person distributing the content. Requiring **user accounts** on platforms can provide an additional disincentive against problematic synthetic content, and comes with personal identification of the creator or distributor such that they can be blocked or banned if their actions are malicious.

3. Advancing Responsible Global Technical Standards for AI Development

Information on “AI nomenclature and terminology”

NFAA

We have introduced the tag **“NFAA”**, meaning “Not For All Audiences”, to our models, datasets, and demos. This tag notifies users that there are situations where people should not be exposed to the content. Examples of inappropriate audiences include children who should not be able to access content their parents prohibit and bystanders who may see or hear content coming from your computer that they do not want to see/hear.

This term can be contrasted with the common term “NSFW” (Not Suitable for Work), which is generally applied to sexual material. We found this term did not suit our needs as it brings with it the implicit assertions that:

- Which content to access should be grounded on your work: This misses the fact that what you view on your screen might actually need to be tempered by where you are physically located, such as a public coffee shop.
- There are no workplaces where sexual material is appropriate: This is not correct in all cases, such as sex work.

This tag (as well as our Code of Conduct and Content Guidelines discussed in platform governance), are grounded on the value of **consent**. As such, the NFAA tag addresses the un-consented sexual material users might encounter on the Hub.

Open Source, Open Science, and the Gradient of Openness

Casting AI as either “open” or “closed” presents a false binary. To contextualize this with respect to Hugging Face, we are an open platform for AI: We take a community-based approach to understanding how concepts like “open source” and “open” are defined and understood. For traditional software, “Open Source” software has a specific formal meaning whose definition is maintained by the Open Source Initiative and much of the code we develop has OSI Approved Licenses. For AI systems, conversations about how to operationalize the values that underpin Open Source software are still very much ongoing; especially given the importance of data in fully understanding how they are designed and how they work. As such, we tend to use terminology like “open science” and “open access.” For AI models, we create processes in light of the trade-offs inherent in sharing different kinds of new technology, and approach our work in terms of a **“gradient of openness”** (also called a “release gradient”) to foster responsible development. Please see Figure 5 and https://huggingface.co/papers/2302.04844 for further information on the gradient, and our section on Human-computer interface design for methods that fall along the gradient.

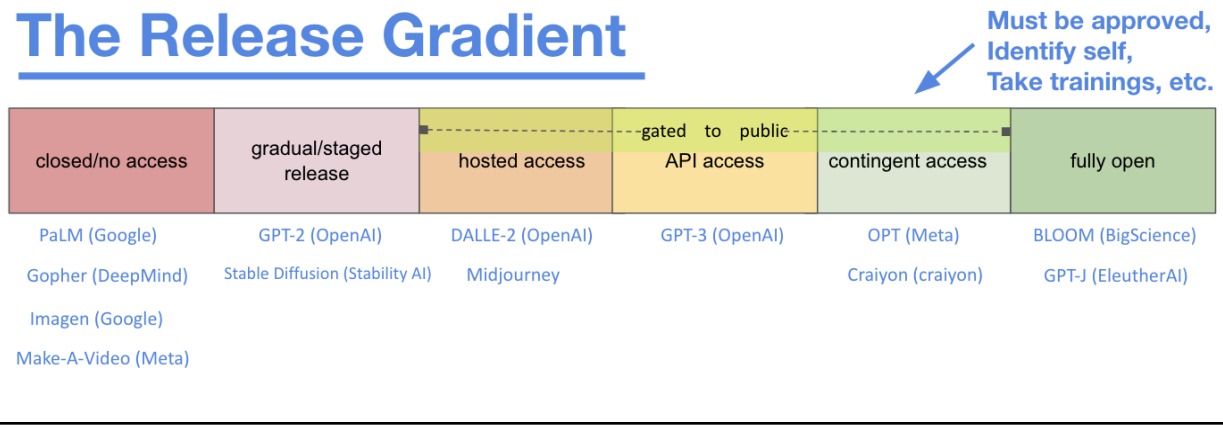

Figure 5: The Gradient of Release
(adapted from Solaiman, 2023, "The Gradient of Generative AI Release: Methods and Considerations")

Information on "collection and use of data"

Addressing excerpt:

> - Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis, as well as inclusivity, fairness, accountability, and representativeness (including non-discrimination, representation of lower resourced languages, and the need for data to reflect freedom of expression) in the collection and use of data;

Relevant information is provided in our sections on Data Governance, Data Governance Roles, Data Measurement, Data Development Roles, and Dataset Documentation.

Given the importance of data in fueling the current progress in Artificial Intelligence, we are convinced in the importance of the **transparency and documentation of this data**, as described in our section on Dataset Documentation. This not only allows for auditing (described further in our section on Auditing) – for instance, allowing data creators such as artists and authors to verify whether any of their data is contained in datasets – but also to better understand what datasets contain and what they're missing (also see our discussion on Data Measurement and Dataset Documentation). While the legality of data usage and aspects such as copyright are still being debated in courtrooms across the country, auditing and documentation are two key contributors towards a more transparent and trustworthy practice of AI, allowing the mitigation of unintended consequences and potential harms of AI-enabled systems.

Privacy

One way to create "privacy" in datasets is to redact private content. In the U.S., a common type of private content is termed "PII", or "Personally Identifying Information". However, **PII detection in text does not always work well**. The state of the art in unstructured text has been defined by

Presidio from Microsoft, which works better for some categories of PII than others, and is U.S.-centric: It doesn't detect types of private information from many other countries (such as other kinds of national identification numbers other than the U.S. social security number).

When applying PII redaction, it is important to analyze the **false positives**, as these may include desired content in data – such as mathematical formulae or dates. A manual audit we ran on Presidio (https://aclanthology.org/2023.trustnlp-1.18/) demonstrated false positives where U.S. bank numbers, social security numbers, and credit cards were confused with ISBN numbers, MLS numbers, article numbers, phone numbers, and miscellaneous manufacturing part numbers.

It is also important to note that there is a **difference between PII detection/identification and verification** – a tool built for one cannot be accurately used for another. The latter focuses on whether a string obeys strict guidelines, such as those defined by the World Wide Web Consortium (W3C). The former provides for the fact that people will write down things like email addresses and websites in ways that don't specifically abide by the defined standards, for example, using both upper and lower case in an email alias.

Information on "Human-computer interface design for AI systems"

Addressing excerpt:

> - Human-computer interface design for AI systems;

Here we briefly discuss the role of **points of friction** where a user would begin interacting with a system, requiring the user to understand the content they are about to engage with before engaging with it.

Gates

Gating is a barrier and mechanism that requires potential users to meet certain requirements in order to access content. This might be simply acknowledging a license. It may also be filling out a form on how they intend to use the system in order to be approved. Or it may go even further, requiring users to take a training course before they are able to use the data, model, or system. For example, https://huggingface.co/StanfordShahLab/clmbr-t-base is a health model that uses Hugging Face gating to require CITI training before anyone can access it.

Modals

These are boxes that users must read. Applied to systems that a user interfaces with directly, such as chatbots, this can be a **point of friction**, where users must read the content in the modal before continuing. This technique was applied for Hugging Chat (https://huggingface.co/chat/; see Figure 6) and is critical to prevent misunderstandings (see, e.g., https://www.theverge.com/2023/5/30/23741996/openai-chatgpt-false-information-misinformation-responsibility).

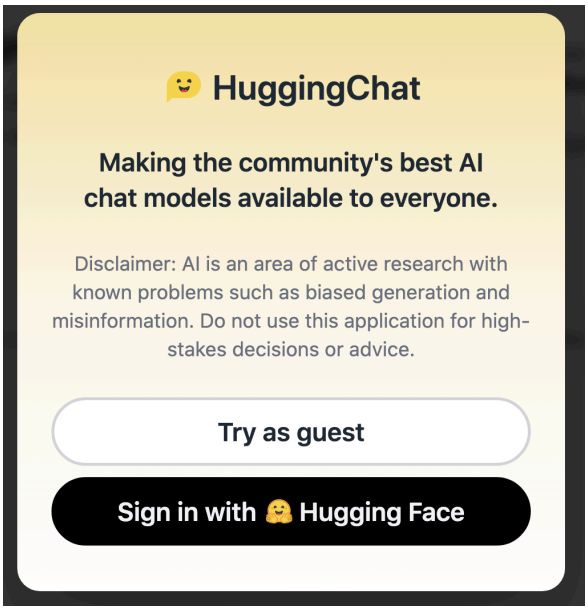

Figure 6. "Modal" that appears before users can interact with HuggingChat (https://huggingface.co/chat/)

Information on "AI-related standards development activities"

Addressing excerpt:

> - Suggestions for AI-related standards development activities, including existing processes to contribute to and gaps in the current standards landscape that could be addressed, and including with reference to particular impacts of AI;

Aside from the work of the International Standards Organization, relevant existing work on standards development for AI includes:

- The Partnership on AI's About ML project (https://partnershiponai.org/workstream/about-ml/), which brings together a diverse range of perspectives to develop, test, and implement machine learning system documentation practices at scale.
- Hugging Face's Model Card Guidebook (https://huggingface.co/docs/hub/model-card-guidebook), which includes a documentation standards landscape analysis (https://huggingface.co/docs/hub/model-card-landscape-analysis) and user studies (https://huggingface.co/docs/hub/model-cards-user-studies) on how people use model cards to document models.

With "reference to particular impacts of AI", please also see the AI Incident Database, https://incidentdatabase.ai.

Appendix

Information on Watermarking for the tracking of generated content from Imatag

As the capabilities of generative AI improve, most detectors of generated content are doomed to fail. Indeed, the objective on which generative AI is trained is to mimic the distribution of real content. Therefore, the artifacts on which most detectors rely are fading away as new AI methods are designed, as was illustrated recently with OpenAI removing its own Chat-GPT detector due to its poor performance.

Watermarking is a method designed to unobtrusively mark content in order to convey additional information, such as authenticity. Within AI, watermarking involves adding machine recognizable patterns to digital content (such as images), where these patterns convey information such as where the content came from and whether it's synthetic. By **proactively** adding watermarks to digital content as it's created, it becomes possible for people and platforms to detect and track digital content as soon as it's shared. New approaches to watermarking, such as those pioneered by Imatag, create watermarks that are imperceptible to humans but can be recognized by algorithms, so as not to disrupt usage of the content. New digital watermarking methods are also designed to be **robust** to common alterations (e.g. compression, cropping, and color changes in the case of images).. AI watermarking solutions are **secure** if malicious attempts to remove, extract, or forge watermarks cannot be done by a third-party unaware of the secret used to protect the solution.

This compromise between imperceptibility, robustness, and security is key to evaluating watermarking systems. As it spreads on the Internet, digital content is often modified multiple times for technical or editorial reasons. For example, it may be recompressed to save bandwidth or cut to optimize rendering on various devices. Current open-source watermarking solutions are not robust enough to these kinds of alterations for practical use.

Two distinct methods are currently being studied for watermarking AI-generated images. The first method involves watermarking the output of the generative AI model **after it's created**, just as can be done with real content. In this context, providing a digital watermarking solution that is fast, robust and absolutely secure is possible, as was demonstrated by Imatag who is delivering such a system for some of the biggest newswires[1] on real images. Current closed systems like DeepMind with SynthID, already include watermarking algorithms in their generative processes[2], but applying watermarks as a post-process in an open system has the strong drawback that it is easily removed by simply commenting out a line of code.

---

[1] https://www.prnewswire.com/news-releases/afp-selects-imatag-to-track-the-uses-of-its-photographs-301550539.html

[2] https://deepmind.google/technologies/synthid/

The second method implements the watermarking process **during the creation** of AI content, which is possible for anyone to apply to open-source AI models like those made available on the Hugging Face platform. Given the impracticality of relying on the community to apply watermarks after image generation, it is essential to distribute AI models already modified to automatically embed watermarks as part of their generation process. This is the reason why Hugging Face has been collaborating with Imatag to provide invisible watermarking for the generative AI community, and why Imatag has been prioritizing watermarking methods that do not alter the quality of model output, while also keeping high watermarking robustness and security levels.

IMATAG's solution for watermarking generated content on Hugging Face stands out as the first independent watermarking technology for AI models. While some AI developers, like DeepMind with SynthID, already apply watermarking algorithms, they typically limit these algorithms to their own models.

To address the problem of scalability, watermark detection algorithms must be extremely reliable and designed to operate at a very low false positive rate. Indeed, with the number of AI generated images reaching 15B/year, one cannot rely on detectors operating at even 0.1% error rate, as studied in the very recent WAVES benchmark, since this would mean accepting 15M images per year that are incorrectly identified as human-made! IMATAG’s solution within HuggingFace focuses on providing certified and calibrated detection probabilities to ensure these requirements are met.

New Key Terminology Introduced in this Document

- **data governance roles**
    - **Data Stewardship Organization**
    - **Data Custodian**
        - **Data Host**
        - **Data Provider**
    - **Data Rights Holders**
    - **Data Modelers**
- **dataset development artifacts**
    - **Dataset Requirements Spec**
    - **Dataset Design Doc**
    - **Dataset Implementation Diary**
    - **Dataset Testing Report**
    - **Dataset Maintenance Plan**
- **disaggregated evaluation**: Evaluation applied to different "slices" of an evaluation dataset, such as subpopulations
- **gates**: A barrier to accessing a dataset, model, or demo, that requires additional procedures from the potential user, such as providing their information, reasons for access, or taking a training.
- **gradient of openness**: Different ways that AI content can be shared publicly, on a spectrum from fully "closed" to fully "open".
- **measuring data**: Like model evaluation, but for data.
- **modal**: A box that users see on a web page while all other page content is deactivated. To return to the main content, the user must engage with the modal by completing an action or by closing it.
- **NFAA**: "Not For All Audiences"
- **points of friction**: User interfacing that requires the user to do something (such as reading a label) before proceeding.
- **prompts, seeking and non-seeking**: Something a user provides as input to an AI system to get it to respond or behave in a certain way. When an AI system returns something the user did not intend, the prompt is "non-seeking" with respect to that content. This is particularly important in the case of explicit content.
- **sociotechnic**: Necessary for identifying different contexts where a model may be used, the different subpopulations that may be differently affected by the model, and how the model may be used in different contexts or with respect to different subpopulations.