HUGGING FACE

Hugging Face Comments on NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

Hugging Face commends the National Institute of Standards and Technology (NIST) on the AI Risk Management Framework (RMF): Generative AI (GAI) Profile, an extensive document identifying categories of risk and action items for actors pertaining to those risks. We offer recommendations to strengthen this document based on our experiences toward democratizing good AI and characterizing risks of systems as an open platform for state-of-the-art (SotA) AI systems. Comments are organized by risk categories with corresponding action items on them. If a section or action 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.

Executive Summary

Hugging Face appreciates this opportunity to provide comments and recommendations to strengthen the NIST AI Risk Management Framework for Generative AI. Our emphasis is on responsible AI development practices, multi-stakeholder collaboration, technical safeguards, and ongoing monitoring across key risk areas. Key focus areas include data provenance, transparency, environmental sustainability, information integrity, security, and mitigating biases and harms. We offer consolidated action recommendations below:

Holistic Approach: Safety, Privacy, Consent, Sustainability by Design

- Adopt a "safety by design" approach, focusing on data provenance, quality, training data curation, and evaluations from early stages.
- Implement data minimization practices and robust consent mechanisms (opt-in/opt-out) for data collection and usage.
- Conduct continuous impact assessments and ensure transparency around data sources, licenses, and processing applied.
- Prioritize environmental impact measurement across the AI lifecycle, including carbon footprint calculations and energy efficiency ratings.

Community Feedback: Diverse Stakeholders

- Foster open science, community engagement, and participation from diverse stakeholders, including civil society groups and impacted communities.
- Encourage public benchmarking efforts, leaderboards, and scrutiny regarding the absence of comprehensive evaluations for new models.
- Leverage community feedback loops, external audits, and inclusive processes to assess potential biases, toxicity, and viewpoint homogenization.

Secure Disclosure and Governance

- Implement structured harm reporting mechanisms and secure disclosure processes for AI incidents and vulnerabilities.
- Mandate that model developers clearly specify the intended scope and capabilities of their models in the documentation.
- Treat any deviations from the specified scope or capabilities as flaws to be transparently reported and addressed.
- Establish standards and guidelines for documenting and disclosing model scope, intent, and detected flaws.

Hugging Face remains committed to contributing to the development of a robust AI Risk Management Framework that upholds safety, ethics, and community-driven innovation.

GAI Risks and Actions

1. CBRN Information

We agree with researchers who have investigated this issue empirically [OpenAI study, RAND study] in that, despite GAI systems' being able to generate novel molecule information and make complex information retrieval easier, the realistic threats of such

attacks materializing with current technology are low, since actors are additionally constrained by material availability and wet lab expertise. **We commend the outlined actions to handle CBRN Information risks, specifically on research, monitoring, and data analysis.** Investment in R&D to conduct more empirical research on realistic threat scenarios (MS-1.1-012), establish processes to regularly monitor model use cases to catch potential unforeseen use cases (GV-4.2-010), and continue to scan both training data and model outputs to limit such information (MG-3.1-007) are critical.

2. Confabulation

Risks of harm related to confabulation are exacerbated by how misleading content is distributed and received, especially in high-risk areas. Recent examples include GAI aided web search, trending topics on social media, and medical use cases. We outline some additional risks: GAI fueled misinformation on the internet can also cause threats to election integrity and democratic processes, specifically eroding trust in information and democratic institutions, and obstructing voting procedures and infrastructures. Additionally, since GAI models are trained on large amounts of data scraped from the internet, confabulated data being released into the internet can lead to future models being trained on this false information, continuing a cycle of confabulation and potentially irrevocably harming information integrity on the internet. Confabulation related risks are not only limited to large scale information integrity but also at a smaller scale when real or made up information is incorrectly attributed to people, which can have real-life consequences.

Addressing confabulation behaviors in generative AI systems requires a multi-pronged approach combining training data transparency, disclosure of sources, and evaluation of a model's propensity to confabulate. Firstly, It is essential for model providers to regularly and openly evaluate models for their propensity to generate incorrect information. Two hallucinations leaderboards on Hugging Face (Hallucinations Leaderboard and HHEM by Vectara Leaderboard), assess select models for their accuracy and reliability. These evaluations help in identifying and mitigating the risks associated with AI-generated content. Secondly, developing tooling for content integrity is crucial in maintaining the trustworthiness of AI systems. Tools that enhance content integrity ensure that generated outputs are reliable, thereby reducing the chances of disseminating false information. Finally, it is important to hold users accountable for the content they create and disseminate using AI models. Terms of services are a tool for accountability ensuring user accounts are responsible for the content produced. Any content downloaded, accessed, or used on the platform is subject to these terms and/or the terms accompanying such content. This accountability is aimed at ensuring that users are mindful of the content they generate and share, fostering a responsible AI usage environment.

We agree with the set of actions laid out in the document, specifically reproducibility (MP-4.1-012), via public benchmarking and leaderboard efforts. We strongly support data provenance action items (MS-2.5-008) efforts, specifically documenting data sources and licenses that can aid in attribution and counter hallucinations. We strongly support both documenting model flaws in a structured approach (MG-4.3-003) and public sharing of such flaws to aid transparency (MG-4.3-004). An action item that we want to highlight for more context is (MS-2.13-001) – domain expertise should not only be used for subjective cases like toxicity detection but also high stakes objective use cases where confabulation might become a threat very fast, such as fact-checking during elections.

3. Dangerous or Violent Recommendations

Addressing dangerous and violent outputs in GAI requires holistic and sociotechnical approaches centering impacted communities, rigorous empirical analysis of dataset and model behaviors, and a fundamental rethinking of what constitutes "safety" from diverse perspectives.

It is critically important to adopt a "safety by design" approach when developing AI systems, rather than solely relying on techniques to detect and block potentially dangerous or violent outputs after the fact. While we support community efforts such as datasets and models created for safety checks we must be cautious about over-relying on them as a complete solution. Red teaming initiatives, such as the Red Teaming Resistance Leaderboard, which specifically aims to measure harm and violence, criminal conduct, unsolicited counsel, and not safe for work (NSFW) prediction, can provide useful signals. Red teaming should ideally be used in conjunction with algorithmic impact assessments, external audits, and public consultation, centering the voices of those most affected by potential harm.

Community-driven approaches that amplify historically marginalized perspectives are essential to safety by design - studies such as examining islamophobic biases in GPT-3 and third-party audits on the sexualization of Asian women in image generation models illustrate the need for thorough and inclusive evaluations. Additionally, we must critically examine the datasets and processes used to create and fine-tune AI models. Unsafe pre-training data can influence models to exhibit racist, sexist, and other concerning behaviors that "scale" with model size. Documentation of dataset creation and carefully examining variables like data age, domain coverage, and other quality metrics at the pretraining stage is a more holistic approach to ensure safety by design. Moreover, we cannot ignore the human costs of annotating and labeling potentially violent content, which has been shown to inflict psychological trauma on data workers. Nor can we neglect the

potential diversity trade-offs – overzealous filtering for "unsafe" content has been shown to hamper the diversity of outputs in both language and image generation models.

We largely align with outlined actions, and specifically want to highlight (GV-4.2-001) which lays out plans for third-party assessments (MS-2.7-016) which lays out guidelines for conducting red teaming, (MP-5.1-006) which puts the focus on expert assessments and human feedback; infrastructure for dialogue, such as Hugging Face community discussions pages, foster feedback and engagement. Additionally, we suggest detailed data quality measurement at the pretraining stage and a holistic safety by design approach (instead of a filter-and-remove approach post-training) that has the potential to remove harmed parties, often minorities, from model outputs and therefore the public sphere.

4. Data Privacy

Protecting personal information and privacy in GAI systems depends largely on responsible training data practices, processing methods, and robust security measures. Similar to our recommendations for confabulation, **it is essential to adopt a holistic set of best practices to ensure privacy by design — such as data minimization, opt-in data collection, dataset transparency, and accountability throughout the AI lifecycle.**

Providers should seek consent and respect the explicit choices of individuals for collecting, processing, and sharing data with external parties, as sensitive data could be leveraged for downstream harm such as security breaches, privacy violations, and adversarial attacks. A key risk arises in third-party hosted systems, where deployed language models can leak private user data like PII or sensitive records inadvertently embedded within training sets, often through proprietary system prompts prepended to user inputs during generation. While screening training data for PII and IP violations is feasible, attempting to check generated outputs for training data attribution is still an open research problem unless there is verbatim copying. Defining and identifying "substantial similarity" across modalities like text, images, and audio is difficult. Rather than focusing guidelines around attributing generated output to training data, more focus should be on implementing robust consent mechanisms and privacy safeguards from the start to prevent personal data leaks and violations during the generation process itself.

Ensuring dataset transparency is vital for safeguarding privacy in generative AI systems, while also acknowledging the privacy tradeoffs involved in determining to whom data transparency reports are disclosed. Publishing details about the sources, licenses, and potential privacy risks of training data allows external auditing and accountability. Providers should document data origins and any processing applied. There are other careful

considerations that should be encouraged at the pretraining stage for privacy, not just via more effective licensing but also carefully considering contextual integrity; generative models should ensure that individuals' data cannot be obtained from contexts in which they do not expect it to appear. Some classical notions of privacy protection, like data sanitization, encryption, anonymization, and differential privacy, can be difficult to translate to the generative paradigm.

An explicit recommendation that we have for this section is data minimization: only the minimum necessary data should be collected, used, and retained for the specified purposes of developing and operating the AI system. This reduces the potential for misuse, unauthorized access, or unintended exposure of personal information. Encouraging data minimization practices is especially important in light of data protection regulations like the EU's General Data Protection Regulation (GDPR) which enshrine data minimization as a key requirement. Recent research has also highlighted the risks of excess data retention, showing the impact of repetition in pretraining data on memorization and content leaks in language models. Data minimization not only reduces privacy risks but can also provide computational efficiency gains by reducing dataset sizes. However, implementation requires careful consideration of the right data practices for each use case to balance privacy with maintaining model performance and generalization ability.

Finally, when it comes to accountability, providers should implement robust, public opt-out mechanisms, or better yet, switch to opt-in mechanisms that allow individuals to restrict the use of their personal data for training generative AI models. Vendors implementing opt-out mechanisms should not resort to dark patterns and provide maximum information to users so that they can provide informed consent. Successful implementations like the policies used by BigCode showcase elements of an effective opt-out program including maintaining public opt-out registries where individuals can submit requests to have their data removed from training sets, providing tooling for data creators/owners to automatically remove or exclude their data from being included, committing to not training future model versions on any opted-out data, and offering redacted search utilities for individuals to check if their information is present while preserving privacy. Implementing a universal opt-out policy is challenging given the diversity of data sources and integrations required for large training sets; however, at a minimum, these practices should be encouraged to uphold privacy standards. These tools should be provided in conjunction with regular audits of model capabilities to test for potential privacy leakage and memorization.

In terms of actions, we highlight establishing transparency policies (GV-1.2-007), continuous privacy impact assessments (GV-4.2-001) – that can be done either by the vendor or in the open via community feedback via tools such as leaderboards along with collaboration with privacy and digital rights organizations, and consent mechanisms

(MS-2.10-006) as effective tools, which can be further specified with our recommendations for a privacy by design approach over a detect and block approach. Additionally, our strong recommendation for data minimization broadly connects to MP-4.1-009, but we recommend either highlighting this point or writing out a separate explicit action for data minimization as a policy.

5. Environmental Risk

Assessing and mitigating the environmental impact of generative AI (GAI) systems is crucial from both a sustainability and ethical standpoint and needs to be done via standardizing metrics, fostering transparency from both model developers and compute providers, and collaborating with diverse stakeholders.

The development, deployment, and ongoing operation of GAI systems can have significant energy demands and greenhouse gas emissions that need to be carefully measured and managed. While we know training and running large models consume a lot of energy, the discussion of environmental risk currently focuses primarily on the training phase [OpenAI, Patterson et al, Anthony et al], but it is essential to broaden the scope and comprehensively highlight the energy impact of deploying and fine-tuning GAI models.

Currently, there is a distinct lack of transparency around how much energy is consumed by models, nor is there enough incentive by model developers to perform these measurements. Hugging Face is working on combating this by pioneering an "Energy Star" rating system for AI models, similar to efficiency ratings for appliances and electronics. Such a standardized rating would quantify the energy consumption and environmental impact of training, fine-tuning, and running inference with different models. This would empower more informed decision-making by AI developers, providers, and users in selecting energy-efficient models aligned with their sustainability goals.

Incorporating environmental impact metrics into widely adopted model documentation standards, like model cards, could further incentivize sustainable practices. Ongoing engagement with impacted communities, civil society groups, and compute providers is crucial to shaping a holistic understanding of environmental impacts beyond carbon emissions. Other factors like water and natural resource usage for both chip manufacturing and data center operations should be considered. A key limitation in the implementation of a robust set of actions in combating the environmental risks of GAI is the lack of standardized measurement variables and uncertainty around relative contributions from different stages (pretraining, fine-tuning, inference) and architectural choices. As a standards body, NIST could play a vital role in establishing measurement guidelines tailored to different AI modalities, which in turn would help public measurement and documentation

efforts. In terms of uncertainty, we strongly advocate for greater transparency from hardware manufacturers and data centers for accurate estimation.

While NIST's recommended actions (MS-2.11-005, MS-2.12-001, MS-2.12-002, MS-2.12-003, MS-2.12-004, MS-2.12-005) broadly align with the need to technically measure environmental impacts across the AI lifecycle, we additionally want to highlight the need to involve grassroots efforts, impacted communities, and climate justice coalitions to complement these efforts. Centralized energy and carbon footprint information via efforts such as the Energy Star project would prevent duplicated measurement efforts.

6. Human-AI Configuration

Concerning risks include emotional entanglement, deceptive capabilities, automation bias, and aversion to AI outputs.

Emotional entanglement occurs when users form emotional bonds with AI systems, which can be used for engagement, monetization, or manipulation purposes. AI systems designed to act in human-like ways can exacerbate this issue, leading to technological addiction, overreliance, and even impairment of value-aided judgment in high-risk areas. For example, Google DeepMind's technical paper on AI assistants warns about the ethical implications of emotionally expressive chatbots, highlighting risks such as privacy invasion and new forms of technological addiction. Studies and real-world experiences further illustrate the potential for users to develop emotional connections with AI, underscoring the need for public education and awareness about these risks.

NIST's recommended actions emphasize the need for disclosing AI involvement in outputs and establishing organizational roles, policies, and procedures for communicating GAI system incidents and performance (GV-1.5-004, GV-2.1, GV-5.1-002). There are several ways in which this can be implemented – we advocate for transparency around models as a system, via greater transparency around the outcomes of different initial system prompts, and we strongly emphasize the importance of documentation such as model cards, which should clearly specify scope and intent of the model. A deviation of model outputs from its clearly defined scope or intent should be considered a flaw in the system and should be transparently reported via public avenues such as the AI Incident Database, AVID, AI Litigation Database, CVE, OECD Incident Monitor, or others.

7. Information Integrity

GAI models can produce realistic content in different modalities often faster than humans can, which can exacerbate threats such as mis- and dis-information and their associated

threats such as false advertisements, election integrity, impersonation, and other harms. **Threat level, believability, and ability to mitigate risk will differ by modality.** In this section, we focus specifically on the harms of intentionally generated content that threaten information integrity. A comprehensive approach to ensuring information integrity should combine technical approaches, governance mechanisms, and collaborative public education.

Technical content verification using watermarking has emerged as a mechanism for ensuring content integrity and authenticity in the face of rapidly advancing generative AI capabilities but can be unreliable and not yet tamperproof. Watermarking provides tools to embed metadata about whether a particular piece of content was generated by a GAI model. There are primarily two approaches to watermarking AI-generated content: embedding watermarks during content creation or applying them post-content production. The former, requiring access to the model, offers robust protection as it is automatically integrated into the generation process. On the other hand, post-production watermarking, while applicable to closed-source models, may not suit all content types, such as text. While a lot of proposed watermarking methods are model specific, model-agnostic universal watermarking is an active area of research.

Among the Hugging Face-provided tooling and hosted collaborative projects for watermarking text, image, and audio modalities are image-specific techniques that complement watermarking to limit non-consensual image manipulation. Some subtly alter images to confound AI algorithms, hindering their ability to process them accurately. Tools like Glaze and Photoguard achieve this by making images imperceptibly different to humans but challenging for AI to interpret. Others, like Nightshade and Fawkes, "poison" images to disrupt AI training assumptions, preventing the generation of fake images. Additionally, signing techniques, exemplified by Truepic's C2PA-standard metadata embedding, link content to its provenance and provide certification for metadata validity and integrate with watermarking to bolster information resilience.

NIST can guide standards for informing or restricting AI usage in areas where trustworthiness and quality are essential (MAP 2.1). For instance, leading AI and computing academic bodies such as ICML, ACM, AAAI, and IEEE all have clear authorship policies that advocate for transparency in disclosing AI-edited content and restrict how they can be used. These guidelines should extend to content providers and media outlets, requiring clear labeling of AI-generated content to prevent the spread of misinformation – a study on the current state of internal GAI policies in newsrooms around the world showed varying levels of scope and enforcement in AI usage for reporting and writing purposes. We support documenting processes such as via governance cards, where model developers can disclose their plans for model use cases and list tools for attribution or detection of

content generated by their model.

Finally, end users can take action to protect themselves from AI-generated mis- and disinformation. **NIST can establish common areas for education and literacy** (GV-6.1-003)**.** Public education initiatives can empower individuals to critically evaluate the content they encounter online. This can be done in person at the community level via local libraries and classes in schools – for example, the Boston Public Library is offering free classes on detecting online misinformation, and several states have implemented mandatory media literacy classes in their education systems. These educational efforts should focus on teaching people how to identify common signs of misinformation, understand the importance of source credibility, and use fact-checking tools effectively. Hugging Face strongly supports these efforts, both via providing free AI education resources that instructors can use in media literacy classes, and via providing online community spaces for journalists to use and audit AI models in more effective and informed journalism. By raising public awareness and fostering critical thinking, misinformation education can play a vital role in protecting information integrity and promoting a more informed society.

8. Information Security

While GAI models offer powerful capabilities, they also introduce new risks and challenges that must be addressed to maintain the integrity and safety of digital systems.

The use of open models in secure systems that do not finetune on user data is a powerful way to protect user data. SafeCoder, developed by ServiceNow and Hugging Face, has been rigorously tested for risks like memorization of proprietary code, can run in a secure, air-gapped manner without internet access, and is fully private by design, avoiding training on user data. In contrast, models that collect user data have faced sophisticated cyberattacks, threatening end users. For instance, Samsung experienced a data leak when proprietary information sent to ChatGPT was memorized and later exposed. Research shows that models like GPT-Neox memorized about 1% of their training set, and similar findings with StarCoder demonstrated its ability to closely reconstruct training samples. This highlights the inherent memorization capabilities of LLMs, posing privacy issues. Models like SafeCoder and BlindChat are set apart due to their design, ensuring privacy and preventing such risks - as has also been verified by external audits. This underscores the **importance of transparency and accountability in AI development, as allowing for community auditing and contributions enables identifying and fixing vulnerabilities** (MS-1.1-002, GV-3.2-002).

As a platform that hosts models, we are mindful of the distribution of malicious code via packaged models (MG-3.1-002). Traditional serialization methods, such as pickle used by

libraries like PyTorch and TensorFlow, pose risks by allowing arbitrary code execution. To mitigate this, Hugging Face introduced a security scanner that employs ClamAV, an open-source antivirus, to detect malware. Pickle import scanning raises warnings about suspicious imports that could lead to arbitrary code execution. We developed safetensors, a secure and efficient format that eliminates the risks associated with pickle. This effort, supported by a collaborative security audit, ensures a safer default format for model storage across various libraries, including transformers. Safe formats like GGUF promote the broader adoption of secure practices. Social validation features also enhance trust in model usage. Similar to platforms like GitHub and npm, users rely on models from reputable sources with positive community engagement. **Social features, reporting mechanisms, and spam detection tools help users identify and avoid potentially harmful models.** Combining safe file formats with trusted sources fosters a secure and trustworthy environment for AI model deployment.

It is important to recognize that AI models do not exist in isolation but are integrated into broader software ecosystems. As such, security considerations should be approached holistically via structured, coordinated, and open harm reporting (MS-2.2-010), drawing insights from established practices and frameworks and tailoring them to GAI threats, such as those used in the management of Common Vulnerabilities and Exposures (CVEs) by the MITRE Corporation. NIST's AI Safety Institute (AISI) could be the counterpart for MITRE for such a coordinated AI flaw reporting initiative. Comprehensive security strategies should encompass a range of actions, including data security and privacy controls, organizational security policies and service-level agreements (SLAs), third-party impact assessments, incident response planning, vendor provenance and contract management, and encryption and secure communication protocols. Red teaming efforts, where ethical hackers simulate real-world attacks to identify vulnerabilities, can be particularly valuable in the context of generative AI models. Events such as DEFCON's engagement with government agencies in conducting such exercises, leverage the collective expertise of its community. This community-driven approach can accelerate the development of effective security solutions tailored to the unique challenges posed by generative AI models, as evidenced by Hugging Face's leaderboarding tools and community efforts in this area (AI Secure LLM Leaderboard, Red Teaming Resistance Leaderboard).

9. Intellectual Property

Balancing the copyright implications of large web-scraped training data with the societal benefits of generative AI requires transparency, rights-holder controls, and responsible deployment practices.

One of the major concerns around generative AI systems is the potential for intellectual property (IP) violations, particularly copyright infringement. GAI systems are trained on massive datasets that contain copyrighted works like books, academic papers, computer code, and more. The training data is often collected from publicly available sources on the internet through techniques like web scraping and crawling [The Pile, The Stack, ROOTS, DoLMA, and LAION]. This raises questions about whether the unauthorized use of such copyrighted material for training AI models constitutes fair use or infringement. As we shared in our response to the U.S. PTO, the early stages of dataset curation and model pre-training should generally align with fair use principles, as the aim is to encode general statistics and patterns across the training data, rather than reproducing or replacing specific copyrighted works. However, certain scenarios, like training on a narrow set of works specifically to create market substitutes, could potentially violate fair use.

A key challenge is defining and identifying "substantial similarity" across different data modalities like text, images, audio, and video. While verbatim copying from training data can potentially be detected, assessing whether a generated output is too similar to a copyrighted work is an open technical problem without clear objective thresholds. This problem is made worse by the lack of transparency about the composition of training datasets, which makes it difficult for copyright holders to assess potential infringement and negotiate terms of use. We recognize that ensuring that fair use remains consistent can put an undue burden on copyright holders, and decrease their ability to advocate for their broad interests. However, a broad licensing requirement for training data could lead to market concentration, excluding smaller actors while providing negligible financial benefits to the original creators.

We support NIST's recommended actions in this area about focusing on transparency requirements and dataset provenance (GV-1.2-007) and applying existing laws (GV-1.1-001) regarding human authorship for determining copyright protection of AI-generated outputs, rather than granting automatic rights to model developers or users. These actions would align with international regulation – categories of stakeholder and other jurisdictions have turned to transparency (EU AI Act) and opt-out (EU CDSM) requirements.

In terms of specific, actionable comments on transparency and dataset provenance, our recommendations are similar to what we suggest in Data Privacy, including implementing robust opt-out or opt-in mechanisms via transparent data governance and indexing tools that allow rights holders to restrict the use of their copyrighted works, data minimization as a model design principle, and safer deployment practices like watermarking, truncating outputs, and filtering for verbatim copyrighted material in model outputs.

10. Obscene, Degrading, and/or Abusive Content

In general, GAI systems built solely for the purpose of creating nonconsensual content should be banned (GV-1.1-005, MP-4.1-007). Addressing the risks of generative AI systems producing obscene or degrading content requires a proactive "safety by design" approach that considers the impact of all development choices from the earliest stages. This begins with the careful curation of the pre-training dataset (Longpre et al., Dolma), as there are risks to the "dark side of scaling" language models on massive, uncurated internet data. Ensuring the training data is free from harmful or biased content is crucial (MS-2.6-002), as relying solely on interventions during fine-tuning or deployment stages (MG-2.2-005, MG-3.2-005) —such as content filters or sensitivity adjustments—has proven insufficient. Instances of AI generating deepfakes and other harmful content despite these measures underscore this inadequacy. **Therefore, a comprehensive safety strategy must be embedded throughout the entire model development lifecycle, from data collection and architecture design to setting clear training objectives and rigorous evaluation metrics.**

For the specific case of risks related to child sexual abuse material (CSAM), the Safety by Design guide developed by Thorn and All Tech is Human provides a valuable and actionable framework. This guide emphasizes the importance of transparency and thorough documentation, ensuring that interventions are based on verified harms and carefully considering potential unintended consequences on privacy, diversity, and biases. Adopting this approach helps mitigate CSAM risks while minimizing negative impacts on marginalized communities or legitimate use cases. The guide strongly recommends involving external stakeholders, particularly those with sociological expertise, in developing and evaluating risk mitigation strategies.

The Safety by Design approach can be adapted to address other forms of objectionable or damaging content beyond CSAM, such as hate speech, misinformation, or explicit violence. As with previous technological transitions like social media, the moderation practices of generative AI systems will likely come under increasing scrutiny from regulations like the European Digital Services Act. This necessitates a detailed, transparent approach to risk mitigation that documents the trade-offs and decisions made. Given the complex social determinations and trade-offs between different notions of safety and inclusiveness, risk mitigation approaches must be developed in conjunction with a diverse range of external stakeholders. This collaborative process should involve sociologists, ethicists, policymakers, and representatives from potentially impacted communities. Such inclusive engagement ensures that interventions are grounded in a comprehensive understanding of societal values and potential consequences, leading to more balanced and effective safety measures.

11. Toxicity, Bias, and Homogenization

Generative AI systems, while powerful, can produce toxic, biased, or homogenized content that propagates harmful stereotypes, hate speech, or ideological viewpoints. This presents significant risks, especially as large language models can generate human-sounding text or realistic content in other modalities like images or audio on virtually any topic at scale. Deploying these models globally is challenging due to distinct cultural values and norms around what constitutes sensitive or offensive content.

Toxicity refers to harmful content like hate speech, violent imagery, explicit adult content, or invasive comments that can cause psychological distress. Evaluating for toxicity is critical but complex, given the subjective and contextual nature of what qualifies as toxic across different cultures, languages, and communities. Solely relying on toxicity detection APIs has limitations, as these tools can exhibit biases, such as over-flagging identity terms or under-detecting coded expressions. Bias in generative AI manifests when models disproportionately represent or marginalize certain demographic groups, ideologies, or perspectives. This can occur through skewed object representations, imbalanced occupational and location biases, or the consistent depiction of harmful stereotypes. Bias often originates from training data that over-represents dominant groups and perspectives. Homogenization occurs when generative models produce mainstream, centrist outputs that conform to dominant cultural norms, failing to capture diverse viewpoints. Outlier perspectives from minority groups may be systematically underrepresented or distorted. This issue is significant even within single countries, where cultural diversity is often inadequately represented. As NIST also identifies, homogenization of viewpoints and performances is a looming threat via model collapse due to the rise of training models on synthetic training data.

To mitigate the risks associated with toxicity, bias, and homogenization in generative AI systems, a multi-pronged approach throughout the AI lifecycle is essential, combining better data practices, continuous model evaluations, and ongoing oversight and accountability (GV-2.1-004, GV-3.2-007, GV-3.2-008, GV-4.1-005, GV-4.2-001, GV-5.1-004, GV-6.2-014, MP-1.1-004). We support challenging assumptions at the design stage and examine the entire context in which a model is situated in a sociotechnical system before pursuing technical methods to counter bias. Implementing proactive data collection and curation practices increases the representation of underrepresented cultures, ideological diversity, and more inclusive perspectives during dataset creation while applying data filters to remove egregiously toxic data. Holistically and continually evaluating generations through participatory processes that engage impacted communities, examining outputs across a range of cultural contexts, languages, and sensitive topics using both automated and human evaluations while being respectful of the potential emotional harms of

annotating such content, and disaggregating evaluation results by subpopulations can ensure comprehensive analysis. Establishing transparency through detailed documentation of training data sources and any processing applied and conducting regular third-party audits to test for biases, toxicity, and cultural normativity are all steps towards a robust oversight framework. While challenging, proactively mitigating risks around toxicity, bias, and homogenization from the data and modeling phases itself is crucial for developing responsible generative AI systems aligned with societal values. Solely relying on detection, filtering and debiasing techniques has limitations.

Hugging Face supports numerous leaderboards and benchmarks to evaluate these aspects and emphasizes community building. A participatory approach is essential because understanding the full scope of potential issues requires diverse perspectives. Hugging Face's commitment to open science and open models allows communities to adapt models to hyperlocal use cases, countering the risks of homogenization. Projects like BLOOM and Aya, which exemplify shared multilingual efforts and a global network of researchers with a shared goal, serve as blueprints for creating inclusive AI systems in the open. Collaborative efforts, such as the CIVICS dataset initiative at Hugging Face, highlight the importance of incorporating diverse voices and perspectives throughout the AI development process. By fostering open science and community engagement, Hugging Face aims to develop generative AI that respects and reflects the richness of global diversity.

12. Value Chain and Component Integration

We agree that third-party components in a system can introduce risks if not vetted properly, and we applaud NIST's systemic considerations. As model developers are the best parties to document models, we support developers of third-party components investing in relevant documentation. The methods by which third-party components are procured and applied should be better documented. While approaches such as leaderboards can help compare models, **more research is needed to provide trusted mechanisms for analyzing and evaluating components such as benchmark datasets.** Documentation should include the developer parties and methodologies for procurement and process documentation (GV-1.5-002). NIST can also provide guidance on the appropriate methods for determining component reliability and tools for comparing components.

We thank NIST for work on this profile and look forward to continuing to provide support.

Submitted by:
Avijit Ghosh, Applied Policy Researcher, Hugging Face
Yacine Jernite, ML and Society Lead, Hugging Face
Irene Solaiman, Head of Global Policy, Hugging Face