CENTER FOR
DEMOCRACY
& TECHNOLOGY

National Institute for Standards and Technology
100 Bureau Drive (Mail Stop 8940)
Gaithersburg, Maryland 20899-2000

February 2, 2024

Re: NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

The Center for Democracy & Technology (CDT) respectfully submits these comments in response to NIST’s Request for Comments regarding NIST's assignments under sections 4.1, 4.5 and 11 of the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. CDT is a nonprofit 501(c)(3) organization that works to advance civil rights and civil liberties in the digital age. Among our priorities, CDT advocates for the responsible and equitable use of technology by government agencies, and promotes the adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems. As of January 2024, CDT is also a member of NIST’s AI Safety Institute Consortium.

Our input focuses on section 1.a of the request for comments and is grounded in our underlying assessment that risk management of generative AI should be considered a component of general AI governance practices rather than presenting a standalone set of risks that demand separate frameworks. To be sure, generative AI may in some cases give rise to different risks than other AI systems. But a common set of concepts, approaches, and infrastructure for AI risk management is needed to lay the foundation for generative AI-specific analysis and intervention, and NIST should avoid indicating otherwise. Too narrowly focusing on generative AI at the expense of paying due attention to the building blocks of AI risk management could encourage piecemeal and fractured organizational approaches that are more likely to overlook key risks and lead to incomplete risk mitigation.

With that understanding in mind, our key points are as follows:

- **NIST should prioritize the development of comprehensive taxonomies of harms and mitigations**, including more research into actual use cases of generative AI models based on user and downstream deployer behavior — but these should be layered into holistics AI risk management practices rather than generative AI-specific practices.
- **NIST should continue to urge developers to prioritize the building of technical and organizational infrastructure necessary for overall AI governance, which can then provide a foundation for generative AI-specific interventions.** These foundational investments remain underdeveloped in general, and over-focusing on generative AI risks

at the expense of a more holistic approach risks redirecting focus from critical structural and organizational efforts.

- **NIST should ensure that AI evaluation and benchmarking methods and metrics, including those specific to generative AI, are rigorous and scientifically grounded, and encourage the development of more effective benchmarks in non-English languages.** While NIST has recognized the importance of carefully defining what is to be measured when it comes to AI risk management, AI actors in the realm of foundation models and generative AI have tended to overlook this need in the interest of rapidly deploying scaled and automated evaluation methods.
- **NIST should prioritize and provide guidance on AI documentation as a necessary prerequisite to effective risk management, governance, and accountability.** Documentation helps to ensure all AI artifacts and use cases are tracked, can be accurately triaged as presenting higher or lower risk, and are subjected to necessary mitigations — and sets a baseline for transparency as well as external auditing and oversight that may be triggered in higher risk circumstances.
- **NIST should continue promoting and developing AI consensus standards through inclusive processes.** As NIST undertakes efforts to drive both its own standards efforts as well as international coordination, it should continue to actively recruit public interest participation and emphasize public interest leadership on technical committees and play a role in coordinating and facilitating public interest consultation including for standards and documents under development.

NIST should prioritize the development of comprehensive taxonomies of harms and mitigations, including more research into actual use cases based on user and downstream deployer behavior.

Fundamentally, generative AI is a subset of AI that should be subject to the same risk management processes as other AI systems. While generative AI may call for additional methodologies and mitigations tuned to specific risks it presents, these practices should be layered into the risk management approaches that apply to AI systems regardless of whether or not they incorporate generative AI. Updated guidance from NIST can help organizations develop taxonomies of potential harms that encompass generative AI by developing an understanding of where the risks of generative AI may resemble or diverge from existing guidance, and to identify mitigations for both known and newer risks.

A particular difference that generative AI highlights is the need for guidance around risk management for general purpose AI systems, which can pose different challenges than AI systems built for use in particular contexts. For example, a narrow system that is purpose-built for resume scanning may pose a high risk of facilitating employment discrimination, but a low risk of encouraging self-harm; a foundation model, however, may pose unknown risk of both when considered in the abstract. Besides inherent challenges in defining and measuring latent biases or inclinations towards surfacing harmful information in generative foundation models, an assessment of such a foundation model will in most cases lack specific visibility into how downstream users may deploy that model or how application developers may fine-tune it for specific purposes — not to mention what affect those customizations may have on the risks posed by a given system.

Risk and harm taxonomies can play an important role not only in helping foresee issues that may need to be addressed, but also providing scaffolding for organizations to proactively compile, structure, and validate interventions that are best suited to minimize those risks. This scaffolding also enables organizations and standards bodies to define the set of circumstances where certain interventions must be integrated prior to system deployment. While not all risks, harms, and relevant mitigations can be defined a priori, this sort of basic risk management infrastructure can significantly reduce ambiguity and prevent obvious risks and mitigations from being overlooked. A structured risk management framework that maps defined risks to a “menu” of mitigations can also ensure that more capacity can be directed to considering what marginal or novel risks particular systems present, and to developing approaches to preventing those harms that aren’t addressed by predefined safeguards.

Synthesis and adoption of harm taxonomies should be coupled with the mapping of resources and approaches that address the harms those taxonomies identify. Taxonomies like the Office of Management and Budget’s list of safety- and rights-impacting uses of AI (with particular attention to updates made to the list over time),[1] human rights risks and harms related to generative AI,[2] and peer-reviewed research identifying risks posed by language models[3] provide helpful starting points to create a harm taxonomy for a given system. That taxonomy should be used as a jumping off point to organize and map the measurement methods and interventions relevant to particular harms — and ultimately to define standards for the interventions that AI developers and deployers should deploy in order to minimize harms to people.

Given the general purpose nature of generative AI systems, practitioners would benefit from guidance around how to prioritize the variety of risks that could be considered; however, we note that prioritization that primarily considers the size of populations likely to be harmed risks leading to systemic underinvestment into harms that already marginalized communities face, and to compounding inattention to risks that disproportionately impact those communities.

[1] Executive Office of the President, Office of Management and Budget. “Proposed Memorandum for the Heads of Executive Departments and Agencies,” November 2023. https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf.

[2] Office of the United Nations High Commissioner for Human Rights, “Taxonomy of Human Rights Risks Connected to Generative AI,” November 30, 2023, https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf.

[3] See e.g., Laura Weidinger et al., “Taxonomy of Risks Posed by Language Models,” *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FaccT ‘22*), June 20, 2022, https://doi.org/10.1145/3531146.3533088.

Consultation with affected communities to identify how risks are evolving will remain a pivotal part of risk management for generative AI just as it is for AI more broadly.[4]

Since generative AI can be used in reasonably stand-alone tools like chatbots as well as be incorporated into downstream applications, we urge NIST to deepen its understanding of both how such systems are used directly by users as well as how they tend to be integrated by downstream deployers who modify those systems for specific tasks and contexts — for example, determining how foundational models are adapted or integrated into uses that directly impact people's lives like hiring or healthcare systems, and whether upstream interventions remain useful or whether they need to be modified or augmented by downstream interventions in such contexts.

NIST should continue to urge developers to prioritize the building of technical and organizational infrastructure necessary for overall AI governance, which can then provide a foundation for generative AI-specific interventions.

In its request for comment, NIST's question regarding what changes AI actors may need to make to current governance practices to manage the risks of generative AI reflects two key assumptions: first, that AI actors have implemented governance practices sufficient to manage the risks of AI in general, and second, that generative AI poses risks that may require different governance practices. Neither of these assumptions should be taken for granted.

AI governance infrastructure remains concerningly underdeveloped. The MIT Sloan School of Management recently found that the majority of businesses are reluctant to agree that they have invested sufficiently in responsible AI,[5] and KPMG found that only 19% of organizations surveyed felt they had sufficient internal expertise to conduct robust AI risk management.[6] Even graduate-level computer science students with significant background in machine learning struggle to identify harms related to particular AI systems,[7] suggesting that organizations that rely on ad hoc assessment of impacts and harms will likely face challenges accurately and consistently managing risk. Early stage companies are even less likely to have teams dedicated to governance and risk management of any kind, let alone responsible AI — but at the same

[4] National Institute of Standards and Technology (NIST). "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," January 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[5] David Kiron and Steven Mills "Is Your Organization Investing Enough in Responsible AI? 'Probably Not,' Says Our Data." MIT Sloan Management Review, October 18, 2023. https://sloanreview.mit.edu/article/is-your-organization-investing-enough-in-responsible-ai-probably-not-says-our-data/.

[6] Bart Van Rompaye, "Responsible AI and the Challenge of AI Risk," KPMG, July 11, 2023, https://kpmg.com/be/en/home/insights/2023/07/lh-responsible-ai-and-the-challenge-of-ai-risk.html.

[7] As discussed in Emily Black et al., "Toward Operationalizing Pipeline-Aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools," *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*, October 30, 2023, https://doi.org/10.1145/3617694.3623259.

time may be particularly eager to integrate AI into their business and deploy AI-driven products to market quickly in order to take advantage of perceived efficiencies and compete with incumbents. In order to address the risks of generative AI, these foundational gaps require sustained attention.

As the AI RMF recognizes, "AI risk management should be integrated and incorporated into broader enterprise risk management strategies and processes,"[8] but CDT's conversations with AI practitioners has indicated that the intense public attention on generative AI may motivate businesses to attend to risks posed by language models while failing to sufficiently invest in the institutional and organizational infrastructure that would support more holistic AI risk management practices. NIST should reiterate whenever possible that generative AI-specific considerations should be grounded in the processes, infrastructure, and practices necessary to facilitate general risk management of artificial intelligence, which can then be built upon to handle risks that may be specific to generative AI or any other particular context.

NIST should ensure that AI evaluation and benchmarking methods and metrics, including those specific to generative AI, are rigorous and scientifically grounded, including encouraging the development of more effective benchmarks in non-English languages.

A critical component of risk management is the suite of methods to identify — and in some cases quantify — risks, since these methods tend to inform when risk mitigation is necessary and are used to determine the efficacy of interventions. However, as NIST articulated in its AI RMF, metrics used to measure AI risk "can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts." We have observed the inclination of AI actors in the realm of foundation models and generative AI to overlook this warning in the interest of rapidly deploying scaled and automated evaluation methods. Accordingly, we urge NIST to ensure this important finding remains centered as it tackles risks related to generative AI.

More rigor is needed in the development of measurement approaches.
Many important properties of generative AI systems such as bias and security are difficult to quantify and evaluate. NIST's recommendations can provide AI practitioners with guidance on how to best assess their measurements by drawing on concepts from measurement science — for example, emphasizing considerations of measurement validity in the selection of evaluations for standards and benchmarks.[9] Such techniques have long been employed in social science

[8] AI RMF 1.0, *supra* note 4

[9] Oskar Van Der Wal et al., "Undesirable Biases in NLP: Addressing Challenges of Measurement," *Journal of Artificial Intelligence Research* 79 (January 10, 2024): 1–40, https://doi.org/10.1613/jair.1.15195.

fields, and more recently in machine learning,[10] to translate abstract concepts that cannot be readily observed into concrete measurements.

In particular, practitioners should be careful when translating, or *operationalizing,* abstract properties into concrete quantitative measurements. Take the concept of fairness as an example. At a conceptual level, practitioners could define fairness in various ways, such as ensuring predictive parity among different demographic groups. To measure fairness within an AI system, practitioners must operationalize their conceptual definition of fairness into mathematical metrics that capture specific, measurable information, such as how frequently two different demographic groups receive positive outcomes from the AI system. Yet a single measurement is unlikely to encompass every facet of fairness that practitioners find important. Furthermore, translating a conceptual definition into an operationalized measurement involves making assumptions about the relationship between the two, which may or may not hold in different conditions. Therefore, practitioners should be deliberate in determining which system properties to evaluate at both the conceptual and operational levels, and employ additional evaluation techniques to assess the chosen measurements' adequacy in capturing the relevant system properties. This challenge will remain relevant when it comes to generative AI, so NIST should be particularly attentive to measurement validity in its AI safety efforts.

Ideally, chosen measurements should be *valid* and *reliable* — that is, tightly related to the overarching goal of a given measurement and stable over time.[11] If measures of risk are invalid or unreliable, mitigations that appear to be responsive to those measurements are unlikely to be effective at solving the core problem.[12] Incorporating multiple measures of the same construct is another way to address validity, since it helps gauge the extent to which a construct that may be multifaceted is captured across a set of different measurements. To date, such efforts — such

[10] Abigail Z. Jacobs and Hanna Wallach, “Measurement and Fairness,” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FaccT ’21)*, March 1, 2021, https://doi.org/10.1145/3442188.3445901; Amy Winecoff, Matthew Sun, Eli Lucherini, and Arvind Narayanan. “Simulation as Experiment: An Empirical Critique of Simulation Research on Recommender Systems” *SimuRec Workshop at the 2021 ACM Conference on Recommender Systems (RecSys ‘21)*, October 2, 2021. https://arxiv.org/abs/2107.14333.

[11] Construct reliability is evident when similar inputs to a measure result in similar outputs (Jacobs and Wallach, *supra* note 10). Lack of reliability may indicate issues either with the measurement itself or underlying problems within the system components. For instance, a significant shift in a previously stable measure of model performance could signal concept drift. Conversely, if a measure of model performance lacks stability from the outset, it is unlikely to offer valuable insights into the system's quality. Although estimating reliability is not as prevalent in machine learning as in inferential statistics, measures of reliability can aid AI teams in diagnosing and monitoring potential problems. Reliability information also aids external parties in assessing the system's suitability to meet their requirements since it provides an idea of worst-case performance rather than merely a point estimate. Reliability might also help practitioners assess whether model performance gains reflect meaningful improvements or rather, changes due to random variation. Some reliability measures that have been proposed include test-retest reliability or internal consistency (e.g., split-half reliability) (Van Der Wal et al., *supra* note 9), but even estimates of system variance such as standard deviation over resampled test sets could be helpful.

[12] Van Der Wal et al., *supra* note 9; Xiting Wang et al., “Evaluating General-Purpose AI with Psychometrics,” *arXiv)*, October 25, 2023, https://doi.org/10.48550/arxiv.2310.16379

as meta-benchmarks like the Holistic Evaluation of Language Models (HELM) framework — have been employed primarily to assess system performance, but may also be useful to evaluate potential risks associated with the system when those risks are difficult to operationalize into single measurements.

Measurement validity is particularly important when it comes to AI safety, since consequential decisions about the launch and use of advanced systems will increasingly be informed by results of these measurements — and improperly scoped measurements could lead to faulty decisions that threaten people's safety, access to opportunity, and well-being. To reinforce requisite attention to measurement validity, NIST should encourage AI practitioners to document both their conceptual definitions of the constructs their measurements are intended to assess, as well as the precise mathematical formulation for their measurements. NIST should carefully review these materials when considering which measurements will best serve NIST's goal of developing guidelines, standards, and best practices for AI safety and security.

One tension in generative AI system assessment arises between adopting standardized system evaluations across generative AI systems and creating bespoke evaluations for individual systems. This tension arises in part because measurements of risk in foundation models may not map directly onto measures of risk in deployed systems and thus may fail to capture real-world harms.[13] On the one hand, standard or shared metrics can help facilitate model measurement, tracking, benchmark-setting, and comparison across actors. Improving both standardization and transparency into how models were developed and evaluated can facilitate better reproducibility of model results.[14] On the other hand, the utility of certain measurements may not hold for systems deployed in different contexts. For example, if two systems have different conceptions of what it means for the system to be fair based on an understanding of how those systems are likely to impact people and communities, a standardized method will fail to capture that context.[15] In its efforts to cultivate the measurement science necessary to accurately and thoughtfully evaluate risks of AI systems including generative AI systems, NIST should encourage AI practitioners designing evaluations to keep in mind both their specific system goals and contexts as well as more general desiderata of generative AI systems across contexts.

[13] Jacob Metcalf et al., "Algorithmic Impact Assessments and Accountability," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FaccT '21)*, March 1, 2021, https://doi.org/10.1145/3442188.3445935. For example, spurious correlations between target labels such as "eyebags" in images containing women may or may not be predictive of misogynistic outputs in downstream systems; *see* Angelina Wang and Olga Russakovsky, "Overwriting Pretrained Bias with Finetuning Data," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 1, 2023, https://doi.org/10.1109/iccv51070.2023.00366 .

[14] Sayash Kapoor and Arvind Narayanan, "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science," *Patterns* 4, no. 9 (September 1, 2023): 100804, https://doi.org/10.1016/j.patter.2023.100804.

[15] Jacobs and Wallach, *supra* note 10.

There is a serious lack of high-quality AI benchmarks in languages other than English, which poses significant barriers to detecting and managing risks of generative AI.

Given the centrality of benchmarks and evaluations to the work of AI safety as discussed in the previous section, it is important to call attention to the fact that the vast majority of benchmarks available to evaluate the performance of language models have been created for the English language.[16] Where non-English AI benchmarks do exist, they often fail to reflect the knowledge and linguistic habits of native language speakers. This gap is the result of a popular, though largely unverified, position among NLP practitioners that large enough models can extrapolate from the rules of one language to learn others, for which it may have trained on far less data (a concept called "cross-lingual transfer").[17] This means that foundation models currently undergo far less rigorous testing in non-English languages compared to English, and those tests model developers do employ lack the cultural context of those languages.[18]

While there is some evidence that the cross-lingual transfer assumption holds for etymologically similar languages,[19] practitioners and standards entities such as NIST ought not assume that

[16] This is in large part because English is the lingua franca of NLP academia — it is the most common language of communication between researchers and the language of the most prestigious academic journals. This creates a virtuous cycle of investment for English and a handful of other high-resource languages, where more raw text data available leads to more clean datasets for evaluation, which leads to more research in those languages, which leads to more benchmarks, publications, and attention for their work. Low-resource languages, those with limited text data available, experience this as a vicious cycle, where a lack of data, tools, and academic prestige similarly fuel one another. See Pratik Joshi et al., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," The 58th Annual Meeting of the Association for Computational Linguistics (ACL), January 1, 2020, https://doi.org/10.18653/v1/2020.acl-main.560; Emily M. Bender, "The #BenderRule: On Naming the Languages We Study and Why It Matters," The Gradient, December 5, 2021, https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/; Gabriel Nicholas and Aliya Bhatia, "Lost in Translation: Large Language Models in Non-English Content Analysis," Center for Democracy and Technology, July 7, 2023, https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/; Alexandre Magueresse, Vincent Carles, and Evan Heetderks, "Low-Resource Languages: A Review of Past Work and Future Challenges," *arXiv*, June 12, 2020, https://doi.org/10.48550/arxiv.2006.07264).

[17] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama, "On the Cross-Lingual Transferability of Monolingual Representations," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, January 1, 2020, https://doi.org/10.18653/v1/2020.acl-main.421; Alexis Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, January 1, 2020, https://doi.org/10.18653/v1/2020.acl-main.747.

[18] OpenAI, for instance, has stated that GPT-4 has state of the art performance in 24 languages — but to test that, they simply used a machine translated version of a popular English-language benchmark, MMLU (OpenAI. "GPT-4 System Card," March 23, 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf). Google tested their Gemini model's non-English performance only on the basic, decontextual tasks of answering math word problems and summarizing text. (Gemini Team Google. "Gemini: A Family of Highly Capable Multimodal Models," December 19, 2023, https://doi.org/10.48550/arXiv.2312.11805).

[19] Juuso Eronen, Michał Ptaszyński, and Fumito Masui, "Zero-Shot Cross-Lingual Transfer Language Selection Using Linguistic Similarity," *Information Processing and Management* 60, no. 3 (May 1, 2023): 103250, https://doi.org/10.1016/j.ipm.2022.103250

large language models have inferred universal rules of language to a degree sufficient to be confident that the results of safety evaluations and mitigation actions will hold across linguistic contexts.

This gap is meaningful for several reasons. First, the inability to accurately detect safety issues in one language may create vulnerabilities for all languages. Research has shown that some English-language AI safety mitigations can be circumvented by prompting a model in a non-English language.[20] A lack of language-specific benchmarks and overreliance on machine translation makes identifying these gaps even more of a challenge. Second, downstream users may take foundation model developers at their word that models work in languages they do not. If downstream application developers believe, as suggested by many AI companies' marketing, that foundation models work equally in all languages, they may deploy systems in languages in which they fail or present higher risk. Even if downstream developers have sufficient resources and interest to test their models in these languages, they may not have the proper benchmarks to do so. Third and equally important, a lack of language-specific AI benchmarks will make language models less accessible to non-English speakers. Without such benchmarks, developers will not even know whether their models are functional in nonmajority languages, nor know whether attempts to improve multilingual functionality have been effective. Inequitable access to language models may lead to other downstream disparate impacts, such as reduced access to information or bias in decision-making systems, which has significant implications for accessibility, human rights, and inclusion of global majority values.[21] These downstream risks may be less apparent to organizations conducting risk management on foundational generative AI models, particularly if they assume that existing multilingual benchmarks are sufficient to identify related risks.

NIST should prioritize and provide guidance on AI documentation as a necessary prerequisite to effective risk management, governance, and accountability.

Effective risk management depends on the existence of clear and robust documentation about the AI systems in question. Documentation helps to ensure all AI artifacts and use cases are tracked, can be accurately triaged as presenting higher or lower risk, and are subjected to necessary mitigations. Documentation is also critical to facilitate internal auditing — that is, validation that required risk management actions have been taken — and sets a baseline for external auditing and oversight that may be triggered in higher risk circumstances.

[20] Yulin Deng et al., "Multilingual Jailbreak Challenges in Large Language Models," arXiv, October 10, 2023, https://doi.org/10.48550/arxiv.2310.06474; "OpenAI's Red Team: The Experts Hired to 'Break' ChatGPT," *Financial Times*, April 14, 2023, https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8.

[21] For a discussion of how generative AI can impose norms and values across global contexts, see Irene Solaiman et al., "Evaluating the Social Impact of Generative AI Systems in Systems and Society," arXiv, June 9, 2023, https://arxiv.org/abs/2306.05949.

While documentation may form the basis of transparency artifacts, the intended purposes of documentation and transparency do not always or fully overlap. Any guidance that NIST develops should differentiate between documentation intended to support AI risk management and governance throughout the generative AI value chain, and transparency measures designed to facilitate deeper understanding among, and enable accountability by, external stakeholders.

For example, documentation designed to help assess the appropriateness of models or datasets for a particular use case may not provide the details or clarity that would facilitate policymakers' efforts to hold corporations responsible for AI failures. Meanwhile, transparency obligations requiring corporations to disclose sources of training data in plain language may be less informative to model developers or deployers seeking to understand how to preserve guardrails from a general purpose model to a specific application but may be useful to enable external scrutiny of data collection practices.

Documentation alone does not guarantee transparency, but good documentation is a prerequisite to transparency: it captures what developers know about a model and creates the opportunity for reflection, informs deliberation on how to best share that information, and highlights where there may be insufficient information to satisfy external transparency needs. Meanwhile, transparency efforts that are not informed and backed up by good documentation risks being misleading or incorrect, or at least impossible to validate. Because robust documentation is critical to facilitating both risk management and building toward transparency, we focus our comments on documentation that supports best practices in generative AI governance. We also address the roles that can or should be played by different AI actors for managing risks and harms of generative AI with respect to documentation, focusing on both foundation model developers and downstream generative AI deployers.

Documentation of generative AI system characteristics

Like traditional machine learning, generative AI systems would benefit from structured documentation to facilitate more effective risk assessment and information sharing among stakeholders. Most AI documentation frameworks, such as datasheets,[22] model cards,[23] and system cards,[24] were developed for traditional uses of machine learning tasks (i.e., classification, regression, and recommendation systems). However, because generative AI systems leverage different training paradigms and are more difficult to evaluate, documentation of these systems can be more complex. The situation is further complicated when downstream deployers belong to different organizations than foundation model developers and accordingly

[22] Timnit Gebru et al., "Datasheets for Datasets," *Communications of the ACM 64*, No. 12 (November 19, 2021): 86–92, https://doi.org/10.1145/3458723.

[23] Margaret Mitchell et al., "Model Cards for Model Reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency (FaccT '19*), January 29, 2019, https://doi.org/10.1145/3287560.3287596

[24] Meta, "System-Level Transparency of Machine Learning," February 22, 2022, https://ai.meta.com/research/publications/system-level-transparency-of-machine-learning/.

have less familiarity and access to details about how the foundation models were created,[25] and may have no obligation or mechanism to report details of any modifications to the foundation model developer.

To date, documentation of generative AI systems exists in a variety of forms — from simplified single page tables with limited information,[26] to relatively detailed documentation following the original model cards specification,[27] to lengthy technical white papers following a format reminiscent of machine learning conferences.[28] While longer discussions of model details in technical papers may be rich in detail, their length and inconsistent structures make it challenging to clearly discern key properties and to compare different models with one another. Standardized documentation will both help internal stakeholders compare information across models, aggregate model documentation into a more navigable model inventory,[29] and if provided to downstream deployers, can help them more easily assess and compare foundation models. It may also help streamline the practice of the documentation process, including through automation where reasonable.[30]

NIST should consider how standard formats such as model cards can serve as a helpful baseline to foundation model developers. Using such structures, foundation model developers could help other stakeholders more easily understand, evaluate, and compare foundation models, even if the exact details contained within documentation may still have some variation. Guidance on which specific information foundation model developers should include within model documentation should avoid overspecification,[31] but be sufficiently detailed to encourage consistency across foundation models. The specific details of what model cards for foundation models should include will vary by deployment context, model type, degree of openness, level of risk, and other factors. Still, each of the key sections of model cards are relevant to

[25] For example, OpenAI makes an API available for developers to access and fine tune OpenAI's foundation models for a variety of downstream applications. These downstream applications could be developed by other large technology companies, startups, or even individual developers, who will not have intimate knowledge of how OpenAI's models were built.

[26] "GROK-1 Model Card by XAI," https://x.ai/model-card/.

[27] "Meta-Llama/Llama-2-7b · Hugging Face," https://huggingface.co/meta-llama/Llama-2-7b.

[28] OpenAI, "GPT-4 System Card."

[29] Patrick Hall, James Curtis, and Parul Pandey, *Machine Learning for High-Risk Applications* (O'Reilly Media, Inc, 2020), https://www.oreilly.com/library/view/machine-learning-for/9781098102425/.

[30] Software engineers in general and AI practitioners specifically often express the desire for more documentation tasks to be automated; see Andrew Forward and Timothy C. Lethbridge, "The Relevance of Software Documentation, Tools and Technologies," *Proceedings of the 2002 ACM Symposium on Document Engineering (DocEng '02)*, November 8, 2002, https://doi.org/10.1145/585058.585065. While not all such tasks should be automated, determining which details may be conducive to automation can ensure that minimum documentation exists for all models and systems and trigger rules-based triaging, allowing more time to be spent on documentation components where ethical and sociotechnical reflection is important.

[31] Vanessa Bracamonte et al., "Effectiveness and Information Quality Perception of an AI Model Card: A Study among Non-Experts," *20th Annual International Conference on Privacy, Security and Trust (PST)*, August 21, 2023, https://ieeexplore.ieee.org/abstract/document/10320197.

foundation models and both their substance and structure can provide important insight into aspects of the model’s development that can inform responsible use.

One of the challenges of documenting generative AI systems is that they are developed in multiple stages, often by different organizations, who modify the system along the way. As a result, documentation of generative AI system characteristics at one stage of development or deployment may be more or less relevant in the next stage. For example, when downstream deployers adapt foundation models using techniques such as prompt engineering or fine-tuning, assessments of foundation model risks often do not translate to the downstream task;[32] in some cases, research suggests that bias from foundation models can be inherited or exacerbated by fine tuning,[33] while in others, problems with a downstream task have been shown to be almost entirely driven by aspects of fine-tuning.[34] The lack of clarity in the relationship between risks that present in foundation models and risks that manifest in downstream tasks raises important questions about what systems characteristics ought to be documented, and by whom.

Given the nature of these questions, though, it seems apparent that several elements of the foundation model system are especially important to document. For example, details about data used to train models (e.g., characteristics of data sources, geographic and linguistic distribution, and any procedures to redact or remove certain training examples) may be important to inform risk management, as will specific procedures — such as block lists, content moderation classifiers, and reinforcement learning with human feedback (RLHF) interventions — that foundation model developers use to prevent models from producing illegal, unsafe, or harmful outputs. While several studies demonstrate that safety guardrails implemented on foundation models can be easily circumvented[35] (suggesting that downstream deployers should not

[32] Some studies show that societal biases in the underlying pre-trained model can propagate to fine-tuned models downstream. See e.g., Wang and Russakovsky, *supra* note 13; Hadi Salman, “When Does Bias Transfer in Transfer Learning?,” *OpenReview*, September 29, 2022, https://openreview.net/forum?id=r7bFgAGRkpL. Others show that foundation model bias and downstream bias are uncorrelated. See e.g., Laura Cabello et al., “Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models,” *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, January 1, 2023, https://doi.org/10.18653/v1/2023.emnlp-main.525; Yang Cao et al., “On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations,” Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), January 1, 2022, https://doi.org/10.18653/v1/2022.acl-short.62; Ryan Steed et al., “Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models,” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), January 1, 2022, https://doi.org/10.18653/v1/2022.acl-long.247)

[33] Wang and Russakovsky, *supra* note 13

[34] Steed et al., *id.*; Cabello et al., *id.*

[35] Adversarial fine-tuning on a limited set of examples can increase model outputs promoting illegal activity, hate speech and harassment, fraud, malware, and other harms. (Xiangyu Qi et al., “Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!,”*arXiv*, October 5, 2023, https://arxiv.org/abs/2310.03693), and can cause leakage of personally identifying information (Xiaoyi Chen et al., “The Janus Interface: How Fine-Tuning in Large Language Models Amplifies the Privacy Risks,” *arXiv*, October 24, 2023, https://arxiv.org/abs/2310.15469). Fine-tuning on seemingly

interpret information about the safety of foundation models as translating to their own downstream tasks), better documentation of the steps involved in the foundation model development pipeline will help deployers and other stakeholders more accurately gauge how robust the foundation model's safeguards may be against modifications made by downstream deployers for specific use cases.

Since generative AI systems rely on a multitude of components beyond a foundation model such as user interfaces, safety classifiers, and prompt engineering interventions, higher-order, system-level documentation can also help both internal and external actors better understand the computational pipeline that produces the foundation model outputs and how steps in this pipeline could affect downstream deployments. Both civil society actors and deep learning experts have noted the utility of such high-level descriptions of systems. The idea of "system cards," for example, built on the documentation frameworks geared toward individual system components like datasets or models, and may be helpful in addressing this gap by capturing how multiple system components interact.[36] For example, high-level narrative or visual depictions of the step-by-step processes that systems employ (e.g. pre-training, reinforcement learning from human feedback, prompt engineering, safety filters) to transform inputs into outputs can inform external stakeholders, including both downstream deployers and academic researchers, about the contours of these systems. Moreover, the act of producing such depictions can help developers themselves identify potential problems in the system's logic.[37]

The idea of system cards has begun to see adoption by AI system developers and deployers,[38] which is encouraging; however, as with other forms of documentation, system-level documentation will be most useful if it follows a standardized format. We encourage NIST to provide guidance to AI practitioners on how structured documentation can include higher-level depictions of the overarching system, including both core foundation model system elements as well as any additional system components that implement safety guardrails.

Documentation of generative AI system evaluations

Given the state of research on the transferability of properties of foundation models to downstream applications and the questions that remain outstanding,[39] foundation model developers should clearly document system characteristics but the evaluations they used to test the system and the results of those evaluations. Developers should also document and disclose

benign datasets can also weaken foundation model risk mitigations, suggesting that downstream generative AI deployers could unintentionally undermine foundation model safety guardrails. These attacks are possible both on open source models as well as models with fine-tuning APIs.

[36] Meta, "System-Level Transparency of Machine Learning," February 22, 2022, https://ai.meta.com/research/publications/system-level-transparency-of-machine-learning/

[37] Michelle Lam et al., "Model Sketching: Centering Concepts in Early-Stage Machine Learning Model Design," *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 19, 2023, https://doi.org/10.1145/3544548.3581290.

[38] OpenAI, "GPT-4 System Card."

[39] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," *arXiv*, August 16, 2021, https://doi.org/10.48550/arxiv.2108.07258.

the evaluation measures of capabilities, limitations, and risks for any downstream tasks that they have actively considered. That way, downstream developers looking to customize or deploy a model in a particular context that has already been contemplated would have some insights into how well the model seems to perform (and the relevant risks that have already been identified and mitigated) in use cases similar to their own. Deployers could then gauge the robustness of the foundation model's safety mitigations to harms that might arise in their specific downstream contexts.

Downstream developers are also likely to employ foundation models in ways that are not anticipated by the foundation model providers. As a result, downstream deployers may identify issues that are not addressed in the foundation model providers' documentation. We recommend that NIST encourage foundation model developers to provide downstream deployers with a mechanism to report additional problems they discover, and to update their public documentation to reflect these issues and any mitigations employed to address them.[40] This is especially important when it comes to policies and technical requirements for tracing and disclosing errors, incidents, or negative impacts.

Human factors in generative AI documentation

Beyond considerations about which properties of generative AI systems or their development process should be documented, human factors considerations are also critical to the success of generative AI system documentation. To achieve their intended impact, documentation frameworks must be usable by the practitioners responsible for producing documentation, and documentation artifacts must be understandable by the stakeholders intended to consume them. Therefore, NIST's recommendations for documentation of generative AI systems should address not only what information documentation artifacts convey but also how they can most effectively convey it to suit the needs of different stakeholders toward the goal of making systems safer.[41]

While human factors research on AI documentation is limited, several studies point in productive directions. First, documentation *creators* are more likely to be more deeply engaged in the documentation process if they understand what goals documentation serves.[42] Likewise,

[40] The OFTEn method for Data Cards may offer a conceptual framework for incorporating feedback from downstream deployers into foundation model documentation. Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson, "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI," *2022 ACM Conference on Fairness, Accountability, and Transparency (FaccT '22)*, June 20, 2022, https://doi.org/10.1145/3531146.3533231.

[41] Q. Vera Liao and J. Vaughan, "AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap," *arXiv*, June 2, 2023, https://doi.org/10.48550/arxiv.2306.01941.

[42] Angelina McMillan-Major, Emily M. Bender, and Batya Friedman, "Data Statements: From Technical Concept to Community Practice," *ACM Journal on Responsible Computing*, May 8, 2023, https://doi.org/10.1145/3594737; Jiyoo Chang, "Improving Documentation in Practice: Our First ABOUT ML Pilot - Partnership on AI," Partnership on AI, October 11, 2022, https://partnershiponai.org/improving-documentation-in-practice-our-first-about-ml-pilot/; Anamaria Crisan et al., "Interactive Model Cards: A Human-Centered Approach to Model Documentation," *2022 ACM*

documentation *users* are more likely to use documentation artifacts productively if they know why these artifacts were created in the first place.[43] Second, documentation creators should understand which stakeholders will likely use the documentation and adapt the information within documentation artifacts accordingly — for example, using different degrees of granularity or technical detail.[44] Third, although practitioners often express a desire to automate documentation,[45] engaging in deliberative reasoning during manual documentation can play an important role in helping practitioners to think more deliberately about the risks of AI systems[46] and produce more human-understandable outputs than fully automated processes.[47] Fully automated documentation of systems is undesirable as it would fail to capture important sociotechnical details; however, some information such as summary statistics of quantitative model characteristics (e.g., dataset size, number of parameters, quantitative performance metrics) may be good targets for automation. Meanwhile, information about human decision-making processes that informed model development will need to be documented manually, with particular care being taken for systems and use cases most likely to present higher or novel risks.[48]

To begin addressing human factors considerations, NIST could encourage foundation model developers and downstream generative AI deployers to incorporate a section at the beginning of their model cards describing why the documentation is useful and the intended audience for that documentation artifact. NIST could also recommend that developers and deployers define key technical terms and avoid jargon and acronyms. Where documentation artifacts are designed for use by those other than the original AI system development team, some research has shown that documentation teams that employ two or more practitioners with different areas of expertise

Conference on Fairness, Accountability, and Transparency (FaccT '22), June 20, 2022, https://doi.org/10.1145/3531146.3533108.

[43] Emily Bender, Batya Friedman, and Angelina McMillan-Major, "A Guide for Writing Data Statements," Tech Policy Lab, October 2021, https://techpolicylab.uw.edu/wp-content/uploads/2021/10/Data_Statements_Guide_V2.pdf.

[44] For example, model cards and related documentation often contain information that is difficult for non-technical practitioners to comprehend and, as a result, may not enable them to address AI risks that are relevant to their work. See e.g., Q. Vera Liao et al., "Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience," *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23*), April 19, 2023, https://doi.org/10.1145/3544548.3580652; Crisan et al., *supra* note 42; Pushkarna et al, *supra* note 40.

[45] Amy Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata," *Proceedings of the ACM on Human-Computer Interaction* 6, No. CSCW2 (November 7, 2022): 1–29, https://doi.org/10.1145/3555760.

[46] Karen Boyd, "Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data," *Proceedings of the ACM on Human-Computer Interaction* 5, No. CSCW2 (October 13, 2021): 1–27, https://doi.org/10.1145/3479582.

[47] McMillan-Major et al, *supra* note 42.

[48] Pushkarna et al, *supra* note 40.

who produce the artifacts collaboratively may lead to more robust documentation practices[49] and as such should be encouraged.

NIST should continue promoting and developing AI consensus standards through inclusive processes.

Best practices around the use of data
As articulated in the AI RMF, privacy is an important characteristic of trustworthy AI. While AI systems may pose heightened privacy risks due to their ability to learn and act on patterns that may otherwise be obscure, generative AI in particular has demonstrated the propensity to memorize personal information, and sometimes reveal that information directly as a system output,[50] and so warrants continued and inclusive conversation to identify and incentivize effective practices to address these and other privacy issues.

The AI RMF notes that while developers should consider tradeoffs with other trustworthy characteristics, “privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment.” In the context of generative AI, these values should still be centered. For example, developers of generative AI systems should collect and process only the data needed to provide the service or product they are offering, rather than vacuuming up all data accessible to them. In particular, developers can actively avoid training on sources of data known to have a significant amount of identifiable information and take steps to exclude particularly sensitive forms of data such as biometric data, which could help prevent such data from being inadvertently revealed by users of the resulting model. More thoughtful review of data sources can facilitate such data minimization (as well as help address other safety risks such as the generation of CSAM material or lowering barriers to access for advanced biological or chemical synthesis techniques), as can developing and deploying automated tools to scrub instances of personally identifiable information from training datasets, such as full names, email addresses, Social Security numbers, and credit card information.[51]

To advance the development of consensus standards grounded in trustworthy AI characteristics like privacy, NIST should encourage the conducting of membership inference attacks (attempts

[49] In one investigation, teams of AI practitioners were asked to produce data documentation using one participant as the documentation “author” and one as the documentation “interviewer.” The interviewer asked questions of the author and recorded their responses, taking care to ask clarifying and follow up questions. McMillan-Major et al, *supra* note 42.
[50] Nicholas Carlini et al., “Extracting Training Data from Large Language Models,” *30th USENIX Security Symposium (USENIX Security 21)*, August, 2021, https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting. .
[51] While useful, questions may still remain about how to handle information that resembles PII or sensitive data types but is not, such as the names of public figures or business phone numbers. Such tools also remain imperfect and will tend to reflect similar gaps as AI models themselves (e.g., performing better on common data formats and in majority languages), and so must be combined with additional efforts to reduce the surface area of privacy risks.

to determine the presence of particular training samples or extract training data from a model) as part of model red-teaming, and the development of technical approaches to protecting against such attacks. NIST should also continue exploring the role of privacy enhancing technologies in the training of AI systems, such as using differential privacy or training on encrypted data, as well as the potential for synthetic data both for training and evaluation of generative AI systems.

Ways to improve inclusivity of stakeholder representation in the standards development process
Standards bodies around the world have prioritized AI-related topics to define baseline practices and promote a trustworthy AI ecosystem, but standards development processes tend to be largely inaccessible to public interest stakeholders who have an interest in ensuring that standards effectively uphold rights and reduce harms to people and communities. Obstacles to such participation include lack of visibility into standards development workstreams and financial burdens of membership and publication fees to access standards documents.[52] NIST could follow the example of the UK, which as part of its National AI Strategy launched an "AI Standards Hub" to help stakeholders navigate and more actively participate in international standardization efforts,[53] either by setting up similar hubs or collaborating with such efforts to encourage broad and interdisciplinary stakeholder participation. As NIST undertakes efforts to drive both its own standards efforts as well as international coordination, it should continue to actively recruit public interest participation, emphasize public interest leadership on technical committees, and play a role in coordinating and facilitating public interest consultation including for standards and documents under development.

***

We appreciate NIST's continued solicitation of feedback from stakeholders and affected communities on these important matters. For additional information, or any inquiries, please contact Miranda Bogen (mbogen@cdt.org), Director of CDT's AI Governance Lab.

[52] For example, ISO standards cost upwards of $200 per publication to access. See e.g. https://www.iso.org/standard/81230.html.
[53] AI Standards Hub, https://aistandardshub.org/the-ai-standards-hub.