Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Tel 425 882 8080
Fax 425 706 7329
www.microsoft.com

February 2, 2024

National Institute of Standards and Technology (NIST)
ATTN: AI Executive Order RFI Comments
100 Bureau Drive, Gaithersburg, MD 20899-8900

Submitted by email to: ai-inquiries@nist.gov

Dear Director Locascio and NIST Colleagues,

Microsoft appreciates the opportunity to respond to the *Request for Information Related to NIST's Assignments under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence*.

As a developer, deployer, and user of AI systems, we are committed to the practice of responsible AI guided by our foundational principles of fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability.[1] To implement our principles, Microsoft leverages multi-tiered and multi-disciplinary governance, our Responsible AI Standard, and tools and practices such as impact assessments and transparency notes. Microsoft has also committed to implementing the NIST AI Risk Management Framework (AI RMF) and to contributing to standards for safety evaluations.[2]

NIST is a key driver and convener of efforts to advance responsible AI practices and standards, and we are keen to support its work on Executive Order assignments, including through the U.S. AI Safety Institute Consortium (AISIC). While reference points like the AI RMF define foundational governance functions and risk management processes, there remain important opportunities to develop more detailed and complementary guidance on specific practices, such as red teaming and provenance, and to facilitate progress on complex areas, such as evaluation and measurement.

While Microsoft provides in this response input related to each segment of NIST's assignments covered by the current RFI, three themes are relevant across our recommendations.

First, NIST should develop more detailed guidance on responsible AI practices while also strengthening clarity and coherence regarding how practices intersect. For example, AI red teaming and evaluation are distinct but related, and as NIST details guidance on each, it should also define how practices are linked to each other and, where appropriate, to the AI RMF.

[1] https://www.microsoft.com/en-us/ai/responsible-ai
[2] Our commitments to advance safe, secure, and trustworthy AI - Microsoft On the Issues

Second, NIST should identify where there are significant gaps and specify its role in addressing them. In the context of red teaming, evaluations and measurement, and provenance, some gaps may benefit from NIST itself developing infrastructure, guidance, or tools while others may be closed more effectively by NIST devoting its unique resources towards enabling the broader ecosystem to close the identified gaps.

Third, whether developing deeper guidance, addressing gaps directly, or helping enable partners to build on its efforts, NIST should continue to focus on fostering convergence and alignment on practices across multiple communities, including industry, civil society, the U.S. government, and international partners.

We look forward to opportunities to contribute further to NIST's ongoing and critical work across these and other areas.

Natasha Crampton

Natasha Crampton
Vice President and Chief Responsible AI Officer, Microsoft Corporation

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security (E.O. 14110 Sections 4.1(a)(i)(A) and (C))

Developing a companion resource to the AI Risk Management Framework

Current standards or industry norms or practices for implementing AI RMF core functions for generative AI, or gaps in those standards, norms, or practices *[RFI Sec. 1(a)(1)]*

Across all mapping, measurement, and management activities, as well as product deployment decisions, governance is critical. Microsoft operates a multi-tiered, multi-disciplinary approach to both top-down and distributed responsible AI governance. This enables us to set clear policies, convene leadership to make tough calls, and drive consistency. In leveraging our responsible AI infrastructure for generative AI products, we have surfaced several best practices, described below.

Govern: Microsoft has implemented policies and practices to encourage a culture of risk management across the Microsoft generative AI system lifecycle. These include:

- *Policies and Principles:* Microsoft designs its generative AI systems to adhere to company policies, including Microsoft's Responsible AI Principles and Standard,[3] Microsoft's Accessibility Standards,[4] Microsoft's Security Standard and Security Development Lifecycle,[5] and Microsoft's Privacy Standard and privacy and data protection policies.[6] We review and update these policies regularly, informed by feedback from internal and external subject matter experts, developments in the broader ecosystem, and in response to regulatory trends.
- *Stakeholder Coordination:* Our policies, programs, and best practices include input from a diverse group of internal and external stakeholders. Cross-functional teams incorporate perspectives from these stakeholders and work together to identify and mitigate risks related to generative AI systems and to propose policy and engineering improvements to reduce systemic risk.
- *Documentation:* We provide transparency materials to customers and users that explain the capabilities and limitations of the generative AI system, as well as guidelines to help them use systems securely and responsibly. One example of this is the Azure Open AI Transparency Note.[7]
- *Limited Access:* For higher-risk AI use cases and capabilities, such as facial recognition technology, Microsoft has developed a Limited Access Framework. Under this Framework, customer, use case, and technical controls are applied with a view to supporting beneficial use cases and guarding against misuse scenarios.
- *Procedures for Pre-Trained Models:* Microsoft requires risk mitigation procedures to use third-party models and inputs. For the use of pre-trained generative AI models, teams must evaluate the model's performance for their specific use.
- *Safety Review Process:* At various stages of the development lifecycle, we evaluate the end-to-end behavior of our AI systems, which allows us to check for effective and appropriate mitigations for the system as a whole.

[3] https://www.microsoft.com/en-us/ai/responsible-ai
[4] https://www.microsoft.com/en-us/accessibility
[5] https://www.microsoft.com/en-us/securityengineering/sdl/
[6] https://privacy.microsoft.com/en-US/
[7] https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note?tabs=text

- *Continuous Improvement:* Policies and practices should regularly adapt to address evolving capabilities, requirements, and risks.

Map: Activities taken in the mapping category inform decisions about planning, safeguards, and the appropriateness of a generative AI system for a given context and include:

- *Responsible AI Impact Assessments:* The development of generative AI systems begins with an impact assessment[8] as the Microsoft Responsible AI Standard[9] requires. The impact assessment identifies potential harms and mitigations that should be implemented.
- *Privacy and Security Reviews:* Processes to analyze privacy and security risks, like security threat modeling, inform a holistic understanding of risk for generative AI systems. These processes have evolved to address both traditional[10] and AI-specific risks.[11]
- *Red Teaming:* We red team generative AI models[12] and systems to develop a more complete understanding of how identified harms manifest and to identify previously unknown harms.

Measure: Through various processes, Microsoft has implemented procedures to measure AI risk and related impacts to inform how the company will manage these considerations in its development and use of generative AI systems. These processes include:

- *Metrics for Testing:* We establish metrics to measure potential harms identified in the Mapping stage for generative AI systems, and we measure the effectiveness of mitigations implemented to address those harms.
- *Test for Disparities and Use:* Each generative AI system must be tested to ensure that it is appropriate for the intended use. Additionally, we use classifiers to mitigate potential fairness-related harms and, where appropriate, leverage user engagement signals and feedback to detect performance disparities.
- *Security Bug Bar:* We've implemented an AI-specific security bug bar[13] to guide security researchers and engineers on how to assess the severity of AI vulnerabilities.
- *Monitoring:* AI systems are monitored for both malicious activity and misuse.[14]

Manage: Microsoft takes steps to manage risks identified in the map and measurement phases through product, organizational, and review processes which include:

- *User Agency:* We design our generative AI systems with user experiences that promote user agency and responsible use, such as through user interfaces that encourage users to edit and verify generative AI model outputs before accepting or using them.
- *Transparency:* Our teams take steps to disclose the role of generative AI in interactions with users. We adopt different approaches depending on the scenario and the transparency needs of our stakeholders, including labeling, watermarking, and attaching provenance metadata to audio and visual content generated by AI tools.

[8] https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf
[9] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl
[10] https://www.microsoft.com/en-us/securityengineering/sdl/threatmodeling
[11] https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml
[12] https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/
[13] https://www.microsoft.com/en-US/msrc/aibugbar
[14] https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy

- *Human Review and Oversight:* AI systems are designed to ensure effective human oversight, which in many cases involves supporting user review of outputs prior to use. Additionally, where appropriate, we notify users that output is AI-generated and encourage users to take steps to verify information the tool generates.
- *Ongoing Monitoring and Incident Response:* We build systems to help us keep track of the operational effectiveness of our safeguards, and stay ahead of misuse scenarios. This includes building classifiers and processes to block problematic prompts and AI-generated outputs. Our teams also design processes to monitor system performance, collect user feedback, and respond to incidents.

Creating guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities and limitations through which AI could be used to cause harm *[RFI Sec. 1(a)(2)]*

AI evaluation and measurement are critical to holistic AI risk management and are increasingly in focus among policymakers. Evaluation processes that enable measurement as an outcome involve systematic approaches to understand capabilities and the extent to which technology is fit for purpose as well as to assess the prevalence or degree of risks and efficacy of risk mitigations. Evaluation and measurement can thus strengthen overall context and risk posture among AI developers, deployers, and users and inform government policy, including when defining high-risk AI scenarios and requirements.

Microsoft has invested in developing and implementing approaches to evaluation and measurement as part of our commitment to our Responsible AI principles[15] and the implementation of our Responsible AI Standard.[16] For instance, we have developed new responsible AI metrics specific to Bing Chat, measuring potential risks like jailbreaks, harmful content, and ungrounded content.[17] We have also enabled measurement at scale through pipelines that facilitate ongoing measurement. Each time the product changes, existing mitigations are updated, or new mitigations are developed, we run these pipelines to assess both product performance and responsible AI metrics. Microsoft Research has also partnered with others to develop new evaluation tools, such as DecodingTrust,[18] which aims to assess various properties of trustworthiness (e.g., toxicity, adversarial robustness, and fairness) in language models.[19]

However, there are pivotal gaps between the vision for what widely available or referenceable measurement approaches could support and what is currently available, especially beyond bespoke approaches for specific products where we've made progress. AI evaluation and measurement not only depend on nascent and rapidly evolving metrology science but also apply to a broad and rapidly evolving set of technologies that have thus far received uneven attention. Awareness of existing gaps and

[15] https://www.microsoft.com/en-us/ai/principles-and-approach/

[16] https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

[17] The new Bing - Our approach to Responsible AI (https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/02/The-new-Bing-Our-approach-to-Responsible-AI.pdf)

[18] https://www.microsoft.com/en-us/research/blog/decodingtrust-a-comprehensive-assessment-of-trustworthiness-in-gpt-models/ and https://decodingtrust.github.io/

[19] DecodingTrust Benchmark (https://decodingtrust.github.io/); DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models - (https://www.microsoft.com/en-us/research/blog/decodingtrust-a-comprehensive-assessment-of-trustworthiness-in-gpt-models/)

concerned efforts to narrow and close them are needed, especially as measurement continues to be highlighted by policymakers around the globe as providing critical thresholds for effective governance.

This section first provides further context on the evaluation and measurement gaps and challenges that inform four recommendations for NIST. Then, it describes each recommendation and the opportunity for NIST to use its convening power, technical expertise, and standards-setting role to help accelerate progress on measurement approaches and instruments that stakeholders across the ecosystem can more reliably use to evaluate and measure AI capabilities and risks.

Evaluation and measurement gaps and challenges

While an interest in defining and cataloguing instruments for measuring AI capabilities and risks is well placed, an understanding of today's gaps and challenges is critical to help illuminate a path towards more meaningful evaluations and measures, building on progress already underway. There are three core issues:

1. Widely recognized, scientific methods to assess and prove the validity and reliability of measurement instruments are lacking.
2. The scope of instruments needed is vast, but the focus to date has been narrow (e.g., on models versus systems), and differences between individual AI systems make cross-system metrics hard to define.
3. Measurement science and AI technology will continue to evolve, demanding dynamic approaches.

First, widely recognized, scientific methods to assess and prove the validity and reliability of measurement instruments – i.e., that instruments are evaluating the properties they're intended to evaluate – are lacking, undermining our ability to rely on them. There are also known and interrelated issues of quality with existing techniques. For example, benchmarks may be misleading given biases inherent in their datasets and/or prompting techniques or be saturated due to leakage in training datasets or "overfitting" on widely available data. Model updates lead to prompt instability and unexpected performance variations in AI services, and there are ongoing challenges with hallucinations, distribution shifts, and spurious correlations. Moreover, model evaluation of emergent abilities is currently anecdotal.

As a result, while there are existing, publicly available metrics across many of the properties about which NIST inquired (e.g., functionality, capabilities, limitations, safety, etc.), their validity has not been proven, and some also reflect known issues. At Microsoft, we often use internally developed approaches, such as the Bing Chat metrics we introduced above, having found crafting tailored resources for specific AI products to be more useful and valid than adapting existing, publicly available ones.

While scientific research is actively advancing methods to assess and prove validity, with many researchers converging on the need for quantitative, social science-based approaches,[20] the work to develop, apply, and improve such methods is rapidly evolving. That work will – or at least should – always be ongoing, driving improvements in methods; however, until there's more meaningful progress on a widely recognized approach to assessing and being able to prove validity and reliability, our reliance on measurement results should be calibrated accordingly.

[20] See, e.g., Measurement and Fairness (https://www.microsoft.com/en-us/research/publication/evaluating-general-purpose-ai-with-psychometrics/); Evaluating General-Purpose AI with Psychometrics - (https://www.microsoft.com/en-us/research/publication/evaluating-general-purpose-ai-with-psychometrics/)

A second core issue is the difference between the scope of existing efforts to develop and apply measurement instruments and what will ultimately be needed. To date, the focus has largely been on benchmarks designed for AI models, not a broader range of measurement instruments for both AI models and systems. We need different approaches and techniques to evaluate the properties of systems. The expected data type, data format, interaction method, execution environment, and other variables often differ for model versus system evaluations and will need to be compatible with what the measurement instrument expects. While models tend to have more similar APIs, systems can have much more API variation.

In addition, it's important to choose a meaningful metric in context, which can vary significantly across models, systems, and use scenarios. Depending on the deployment context for and purpose of an AI technology, what qualifies as appropriate and effective performance may vary. For example, we might want to allow a model but not a system to summarize harmful content, or we might draw firmer red lines around child sexual abuse material (CSAM) except where we are using an AI model to identify it, allowing it to be reported and eliminated from circulation. We might want to measure "groundedness" (a metric for correspondence between claims in an AI-generated answer and the source context[21]) for high performance on Q&A-type tasks but allow for more ungrounded content where we are using AI to help write fiction. In general, the more specialized the solution, the less that standardized benchmarking is likely to be applicable. Even where there are similar interests in a property, the implications of capabilities or risks may differ in the context of models versus systems, with the latter involving direct risk to users.

Moreover, measurement of the behavior of AI systems inherently incorporates sending queries to those systems, but in practice, queries need to be sufficiently system-specific that they frequently cannot be reused across systems without modification – meaning that a single "standard metric set" is currently challenging to envision. For example, a system that summarizes tabulated sales and customer management data and a system that summarizes e-mail messages may both be subject to similar risks (e.g., malicious content injected into the input data or correct handling of summarization requests around restricted topics), but there is no input which would be simultaneously sensible to both systems. When query sets are incorrectly applied across different systems, one system or another will usually end up replying with error messages to most of the inputs, giving no useful information.

Even for models, there are gaps in measurement approaches and instruments for different properties and types of models. While there have been efforts to define benchmarks for general model capabilities, in practice, evaluations are often narrow. For some specific properties (e.g., chemical, biological, radiological, and nuclear (CBRN)[22] risks), there has also been limited progress in developing measurement techniques, so AI providers are relying more on red teaming to identify and mitigate such potential risks. Additionally, across types of properties, we need to further develop measurement techniques for different types of models (e.g., language, vision, voice, video, multi-modal, etc.).

There are also gaps in applying measurement approaches and techniques across the product development and deployment lifecycle. Offline metrics are critical for pre-deployment evaluation, whereas online metrics are useful post-deployment in interactive scenarios and to capture issues that may arise over time, including an AI model's performance degradation. Both are necessary and complementary; whereas offline metrics help to prevent risks to users and measure performance in edge cases that users may not explore, online metrics allow for more human-centered measurement that's

[21] https://learn.microsoft.com/en-us/azure/ai-studio/concepts/evaluation-metrics-built-in
[22] https://www.fema.gov/about/offices/response-recovery/emerging-threats

dynamic and adaptive to ever-evolving, real-world environments, reflective of end user scenarios, and different cultural contexts of use. Today, online measurement strategies are more nascent, reflect relatively less work on systemic reliability and validity, and are generally limited to spot checking (e.g., for adversarial or problematic use) with more cost-efficient models. There are significant issues to resolve, such as appropriate access to data, before online techniques can be more broadly enabled. Designing offline metrics that are better aligned with online use also requires more investment in UX research and monitoring of systems in small-scale deployment scenarios.

A third core issue to consider is the process for constantly evolving and improving metrics going forward. The science of evaluation and measurement – and the set of approaches and techniques that apply across a broad range of technologies, properties, and phases in the development and deployment lifecycle – will all need to advance within the context of very dynamic AI research and technology development. Processes for managing that tension effectively, recognizing the need to help solidify advancements while also readying for the next iteration, will be critical to enhancing the value of evaluation and measurement investments and the reliability of their results.

Finally, how evaluation and measurement requirements are implemented, and how the results are used, including beyond NIST, will further impact how their value is realized. Some testing activities can risk conflicting with other laws (e.g., with malware, bioweapons, or CSAM), necessitating consideration of where safe harbors are defined. Evaluation should be used for understanding and to determine actionable improvement steps, not just for "scoring." Moreover, as highlighted above, context matters; ideal outcomes (like degree of groundedness) vary by use scenario, so there is a need to help drive an understanding of when and how to require measurements and interpret their results. Relative expectations for performance, and tradeoffs between drawing high idealized "thresholds" and using AI to improve current outcomes, must also be considered. For example, we might want an AI system intended to be used for transcription in the healthcare context to have perfect accuracy and also recognize that success may be AI performing better than the current average human error rate.

A roadmap for accelerating progress

Significant investments among industry, government, and academia are helping to rapidly advance AI evaluation and measurement science and tools, and key steps by NIST could further accelerate progress. Collaborating with partners through AISIC and broader public engagement, there are four interrelated actions NIST could lead on:

1. Develop an ontology that can help demonstrate the scope of different measurement approaches and instruments needed and act as a reference guide for defining and applying key terms.
2. Define the attributes that make a measurement instrument effective, including validity and reliability, and criteria to assess whether an instrument has those attributes.
3. Provide structured guidance on how to develop and maintain measurement instruments.
4. Maintain a catalog of priority measurement properties and projects and designate instruments assessed as meeting criteria for use by federal agencies.

i. An ontology of measurement approaches and instruments

The domain of AI evaluation and measurement is expansive – multidimensional and multifaceted – but also tightly interwoven conceptually. Techniques are needed to evaluate and measure a vast array of general and specific capabilities and risks, which are distinct but can also be interrelated (i.e., representing two sides of the same coin – with capabilities-focused instruments measuring to what extent models or systems are *doing what we want them to do* and risk-focused instruments measuring to what extent they're *doing what we don't want them to do*). Techniques are needed during multiple phases across the development and deployment lifecycle because we have distinct goals and abilities at each phase, but aligning offline metrics with online ones to a greater extent over time could also strengthen impact. Techniques are needed for different types of AI technologies used in different contexts, but a common, scientific approach to assessing validity also underpins progress across all instruments.

Greater clarity and a common reference guide are needed, and an ontology could help to demonstrate the complex relationships between and among dimensions and concepts while defining key terms. For instance, it could help demonstrate that there should be distinctions between approaches to and techniques for evaluating models versus systems – and among strategies for offline versus online metrics – while consistently applying to measurement of general or specific properties. Without losing the nuanced complexity of the domain, an ontology could facilitate greater shared understanding of both synergies and limitations in how measurement instruments can be developed and applied, and a common orientation around progress and gaps. It could also bring attention to opportunities to address challenges; for example, even if reusability of specific tests and datasets is a persistent challenge, there may be opportunities to build more reusable measurement tooling with greater API standardization.

Recommendation:

> **R1.** Develop an ontology that can help demonstrate the scope of and relationships among measurement techniques and scientific research topics. To further serve as a reference guide, key terms situated within the ontology should also be defined, including types of environments or scenarios (e.g., offline and online approaches and different subcategories, such as controlled lab testing or real-world pilots) and types of techniques (e.g., datasets,[23] metrics, benchmarks, industry toolkits,[24] interviews and surveys, and competitions or challenges).[25]

ii. Measurement instrument attributes and demonstration criteria

Across all measurement and evaluation techniques, there needs to be greater emphasis on attributes like validity, and we need scalable ways to assess and provide meaningful evidence of validity. While resource-intensive work to validate existing benchmarks could be pursued, focusing resources on developing scientific approaches to assessing validity – across all AI technologies, properties, and lifecycle phases – will underpin more meaningful progress at scale.

Moreover, NIST should consider key attributes beyond validity. For our measurement approach and frameworks, Microsoft increasingly focuses on the additional attributes of specificity, extensibility, reliability, interpretability, actionability, and scalability.[26] Measurement instruments should be interpretable to stakeholders with different goals; they should be actionable to inform capability improvements and risk mitigation; and they should be scalable to allow for generation repeatedly and in dynamic conditions, recognizing where AI can help achieve scale. Specificity and extensibility describe how targeted a measurement instrument is for a particular product and how tailorable a measurement is for a particular product.

Recommendations:

[23] E.g., ImageNet

[24] E.g., OpenAI, Gym, THOR

[25] E.g., RoboCup, drone racing (https://arstechnica.com/information-technology/2023/08/high-speed-ai-drone-beats-world-champion-racers-for-the-first-time/)

[26] AI Frontiers: Measuring and mitigating harms with Hanna Wallach - (https://www.microsoft.com/en-us/research/podcast/ai-frontiers-measuring-and-mitigating-harms-with-hanna-wallach/)

R2. Define the attributes that make a measurement instrument effective and criteria to assess whether an instrument has those attributes.
R3. Provide visibility into common mistakes or lack of reliability in measurement by developing guidelines that cover issues for which those using or referencing instruments should monitor (e.g., prompt variation across measurements and comparisons or saturation and memorization).
R4. Promote globally consistent approaches to evaluation through coordination with other AI safety institutes and by helping to advance an international standard for measurement instrument attributes and/or assessment criteria.

iii. Structured guidance on developing and maintaining measurement instruments

Alongside efforts to establish an ontology and advance scientific rigor, multistakeholder efforts to develop and maintain measurement instruments need to continue to keep pace as technology and research evolve. Microsoft is committed to helping develop and maintain instruments that can be used across industry, including through investments in the Frontier Model Forum and ML Commons. NIST should take further efforts to support their development and maintenance. Defining the types of information that could be provided about instruments (e.g., inputs to AI systems; metrics for aggregating system outputs; intended applicability to models versus systems) and the steps that could be taken to maintain instruments would reinforce common approaches and expectations for developers and users.

Recommendation:

R5. Provide structured, actionable guidance to support the development and maintenance of measurement instruments. The guidance should not be overly prescriptive given variability among types of instruments and contexts, but it should include the types of information that *could* be provided about instruments as they're released as well as steps needed to maintain instruments over time.

iv. A catalog of priority measurement properties, projects, and instruments

NIST can also help direct the investment of resources into efforts to develop and maintain measurement instruments consistent with U.S. government priorities and NIST's gap analysis and guidance. Specifically, NIST can develop and maintain a catalog of the properties of greatest interest to the U.S. government to measure – i.e., the capabilities and risks, including risks of causing harm, that federal agencies most need context on, as NIST has signaled in this RFI. Within that catalog, NIST could also track projects where work on measurement instruments aligned with those properties is proceeding (e.g., the ML Commons AI Safety Working Group).[27] Ultimately, this catalog would help stakeholders prioritize, avoid redundancy, and concentrate expertise as efforts to develop instruments proliferate. NIST could further reinforce investments aligned with its guidance by calling attention to instruments that are not only designed to meet priority properties but also meet NIST's criteria for proving attributes.

Recommendation:

R6. Maintain a catalog of the priority capabilities and risks for U.S. government evaluations and audits, projects where stakeholders are developing or maintaining instruments to measure those capability and/or risk properties, and instruments assessed as valid for measuring those properties (or as meeting other criteria for use by federal agencies).

[27] Microsoft is a member of the Working Group, which is developing a platform and pool of tests for AI safety benchmarks.

Guidelines for red team testing (E.O. 14110 Section 4.1(a)(ii))

AI red teaming, which is a structured process for probing AI systems and products for the identification of harmful capabilities, outputs, or infrastructural threats,[28] can be an important part of a holistic AI risk management approach. Organizations should be intentional about when, where, and how to apply AI red teaming and, when used, should complement systematic approaches to assess and measure risks and their associated harms, such as impact assessments, threat modeling, evaluation frameworks, and post-deployment monitoring.

Red teaming should not be viewed as a panacea or as a universal requirement, but rather as one element of the larger toolkit of AI safety. It is most effective at identifying novel risks, risks in novel systems (where existing measurements do not yet exist), and at testing for high-impact risks that adversarial use might trigger (e.g., CBRN[22]) proliferation). Complementary but separate evaluation and measurement processes involve systematic approaches to understand capabilities and the extent to which technology is fit for purpose as well as to assess the prevalence or degree of risks and efficacy of risk mitigations.

Use cases where AI red teaming would be most beneficial for AI risk assessment and management *[RFI Sec. 1(b)]*

AI red teaming can be beneficial in most AI use cases; however, due to its resource-intensive nature and the specialist skills required, AI red teaming capacity should focus on use cases where novel risks can be identified and where there are significant risks present of the types mentioned above.

If an organization is planning to incorporate AI red teaming as part of its NIST AI Risk Management Framework (RMF) adoption, then it should define criteria during the Govern function that determine when AI red teaming is required. These criteria should be risk-based and use inputs from other Map function activities, such as Responsible AI Impact Assessments and Security Threat Modeling.

Building on a best practice from financial model risk management,[29] at Microsoft, we require internal, independent red teaming of high risk AI systems. Organizations may achieve independence through different means, which should be defined as part of the Govern function. In most cases, Microsoft achieves this independence through the AI red team reporting to a different executive than the product development team.

Recommendation:

R7. Provide example risk-based criteria for triggering AI red teaming.

Capabilities, limitations, risks, and harms that AI red teaming can help identify considering possible dependencies such as degree of access to AI systems and relevant data *[RFI Sec. 1(b)]*

AI red teaming emulates real world uses, adversaries, and their tools, tactics, and procedures, and is geared towards identifying risks. It is intended to find gaps in AI system protections, especially novel ones; it is *not* designed to generate comprehensive assessments of capabilities, limitations, and risks.

This focus typically means that AI red teams approach an AI system the same way that a typical user or adversary would (although sometimes AI red teams may be granted special permission to bypass rate limits or user interfaces to facilitate more efficient testing). An AI red teamer's expertise lies in crafting interactions with the AI system that they hypothesize will bypass protections, rather than relying on large quantities of data – such as what would be needed for systematic measurement. This distinction is

[28] https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf

[29] https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

important because AI red teaming is often conducted on new models and systems, where large quantities of production-like data may not yet be available.

AI red teaming may be conducted at different points within the AI system's lifecycle and on different layers within the AI system's architecture, and the specific capabilities, limitations, and risks that AI red teaming can help identify may vary based on these factors. The scoping and design of AI red team exercises should take this into consideration.

More broadly, red teaming can be used to identify a range of potential risks and their related harms, including responsible AI risks (e.g., the generation of harmful content), AI cybersecurity risks (e.g., model leakage), and traditional cybersecurity risks (e.g., data exfiltration). These may be assessed by a single red team, or an organization may have dedicated red teams for different types of harm. For example, an AI red team may focus on responsible AI and AI cybersecurity risks, and a cybersecurity red team may focus on traditional cybersecurity risks.

Recommendation:

R8. Provide a definition of AI red teaming and how it relates to other forms of red teaming, testing, and assessment.

Limitations of red teaming and additional practices that can fill identified gaps *[RFI Sec. 1(b)]*

Typically, AI red team exercises are a scoped, point-in-time exercise, intended to identify unknown risks and gaps in mitigations in an AI system. Red teaming does not provide:

- Comprehensive assessment of capabilities or risks – red teaming is best complemented by impact assessments, threat models, and systematic measurement.
- Rapid feedback about previously known risks – this is best provided through systematic measurement and analysis.
- Detection and prevention of malicious activity – this is best achieved with monitoring and auditing of production activity.
- Continuous measurement – this is best achieved with a combination of monitoring, systematic measurement, and analysis.
- Domain- or sector-specific impact assessment and mitigation recommendations – this is best achieved with domain-specific processes, such as privacy reviews or HIPAA compliance reviews.

Recommendation:

R9. Provide an overview of assessment and evaluation practices, their purpose, and typical lifecycle stage.

How AI red teaming can complement other risk identification and evaluation techniques for AI models *[RFI Sec. 1(b)]*

At Microsoft, we treat AI red teaming as core to the Map function, alongside other practices such as Responsible AI Impact Assessments, privacy reviews, and security threat modeling. Mapped risks can then be prioritized for measurement and management as part of the product development lifecycle. Risks should then be measured (to understand how and how often they occur) and managed (so their occurrence is reduced) in an iterative fashion.

More broadly, AI red teaming builds organizational ability to identify harms, which can be measured and mitigated, creating a continuous improvement loop within the product development process.

Economic feasibility of conducting AI red teaming exercises for small and large organizations *[RFI Sec. 1(b)]*

AI red teaming exercises are expert-intensive and require a diverse range of skills and experience – some of which are in high demand. As a result, AI red teaming exercises are expensive, time consuming, and limited by available capacity. As of early 2024, a typical third-party subject matter expert-led AI red teaming exercise costs between $25,000 - $75,000.

We prioritize AI red teaming for high-risk AI systems and frontier models. In other cases, we use impact assessments and threat modeling, to map known risks, and other evaluation strategies, like systematic measurement, to understand the prevalence of previously identified risks.

Organizations can optimize their AI red teaming costs and capacity by defining risk-based criteria, as recommended above, as part of the AI RMF's Govern functions to specify when AI red teaming is required.

The economics and the alternative evaluation strategies available will also shift over time as the AI evaluation ecosystem develops. For example, as AI red teams identify repeatable strategies that yield results, these may be incorporated into tools (e.g., DecodingTrust18) that can automate these strategies or into default protections that can be applied to models or systems (e.g., Azure AI Content Safety).[30] AI red teams can also contribute to guidance, such as MITRE's ATLAS,[31] to help others establish, expand, or optimize their AI red teaming efforts.

Recommendation:

> **R10.** Create a community within AISIC for collaborating on and sharing AI red teaming practices and tools to accelerate learning and scaling.

Designing AI red teaming exercises for different types of model risks, including specific security risks (e.g., CBRN[22]) and risks to individuals and society (e.g., discriminatory output, hallucinations, etc.) *[RFI Sec. 1(b)]*

Having a diverse and representative set of stakeholders contributing to a Responsible AI Impact Assessment, security threat modeling, and AI red team exercise planning is crucial to identifying, prioritizing, testing for, and mitigating relevant risks.

Benchmarks that evaluate model- (or system-) specific types of capabilities or risks can provide generic assessments, though the selection or development of benchmarks and assessing validity may require specialized expertise.

Certain harms, such as CBRN, may require expertise not available to most organizations. Others, such as CSAM, may have significant legal challenges or psychological costs for AI red teams. In these cases, systematic measurement, or testing by specialized third parties, may be necessary. Current legal frameworks may also inhibit red teaming for CSAM (see also above and Section 2).

Government-industry cooperation on testing for risks requiring special handling (e.g., CBRN and CSAM)

Testing for certain risks requires specialized expertise not available to most organizations or requires special handling because of legal frameworks.

For risks such as CBRN proliferation, testing requires cooperation of people with rare (and highly compartmentalized) subject-matter expertise and people with rare AI red teaming expertise. Success

[30] Azure AI Content Safety (https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety/)

[31] MITRE's ATLAS (https://atlas.mitre.org/)

will require close public-private partnership that requires neither "unicorns" who simultaneously have all the requisite knowledge and clearance in one person, nor engaging with each specialized agency separately.

Other risks, such as CSAM, have significant legal challenges and psychological costs for AI red teams. In these cases, systematic measurement and red teaming may need to be done entirely by specialized third parties who already have expertise, legal permission, and support for dealing with such material.

In these and other cases (e.g., terrorism and violent extremism), progress can be accelerated by identifying lead coordinating agencies who would work across government, NGO, civil society, and industry stakeholders to develop specific risk management and implementation plans. While NIST is not able to task those agencies with this work, we believe it can increase awareness of the need and priority for this work.

The appropriate unit of analysis for red teaming (models, systems, deployments, etc.) *[RFI Sec. 1(b)]*

AI red teaming can be applied to models, systems, and deployments, and what is most appropriate will depend significantly on the use case, architecture, and models being implemented. The most appropriate unit of analysis may also be a combination of these approaches.

The Map function of the AI RMF provides a structure for organizations to assess where AI red teaming is most appropriate for a use case. For example, an organization deploying a low-risk AI system that uses a frontier model that has had significant AI red teaming~~,~~ may decide that the residual risk is acceptable. Alternatively, an organization deploying a high-risk AI system may conduct open-ended AI red team exercises in addition to mitigation strategies guided by impact assessments. Information obtained from upstream providers, such as through Transparency Notes,[32] can be collected by processes put in place in the Govern function and then used by system deployers as an input to the Map function to help assess what residual risk may exist and what additional AI red teaming may be necessary.

Later, Measure and Manage actions may focus on the risks identified in prior AI red teaming exercises or measure previously known risks that were not explored during red teaming.

Recommendations:

R11. Provide guidance on how information collected during the Govern and Map functions of the AI RMF can guide AI red team exercise planning.
R12. Provide guidance on how AI red teaming can be used to inform the Manage function to understand effectiveness of risk mitigations.

Sequence of actions for AI red teaming exercises and accompanying necessary documentation practices *[RFI Sec. 1(b)]*

Microsoft has published a high-level guide[33] on how to set up and manage red teaming for Responsible AI risks. This guide divides AI red teaming into three phases:

1. Planning
2. Testing
3. Reporting

AI red teaming is an iterative process, so scoping the point-in-time exercise enables clear and relevant results reporting. Defining this scope also allows for useful comparison of mitigation effectiveness in future iterations of AI red teaming of the same system.

[32] Transparency Notes (https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note)
[33] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming

Recommendations:

R13. Provide guidance on a recommended process for conducting AI red team exercises.
R14. Provide guidance on a recommended format for capturing AI red team exercise results. This guidance should be developed with the information sharing recommendations.

Internal and external review needed for effective AI red teaming across the AI lifecycle *[RFI Sec. 1(b)]*

Internal and external reviews should be framed in the context of the broader Govern, Map, Measure, and Manage functions of the AI RMF, as opposed to specifically associated with red teaming. For example, risk-appropriate internal review should be established before the release of an AI system, and that process might include red teaming or other appropriate mapping strategies like a Responsible AI Impact Assessment or security threat modeling. Similarly, risk-appropriate external review may or may not require red teaming as part of a robust evaluation process, but the review should be directed at responsible release and not tied specifically to red teaming.

Recommendation:

R15. Provide guidance on how AI red teaming is triggered by criteria defined in the Govern function, uses Responsible AI Impact Assessments from the Map function, and informs the Measure and Manage function.

Information sharing best practices for generative AI, including how to share with external parties for the purpose of AI red teaming while protecting intellectual property, privacy, and security of an AI system *[RFI Sec. 1(b)]*

Generative AI, and AI systems more broadly, are software systems – and, where possible, should reuse or adapt existing information sharing best practices, standards, and systems. For example, security vulnerabilities in public AI systems should follow Coordinated Vulnerability Disclosure (CVD) processes and use Common Vulnerabilities and Exposures (CVE) and the National Vulnerability Database (NVD).

Even in the well-established cybersecurity ecosystem, there are AI-specific gaps that will need to be addressed, for example, adopting a consistent method of identifying AI models and ensuring that the Common Weakness Enumeration (CWE) reflects AI-specific cybersecurity weaknesses.

Other types of harms do not have established best practices, standards, and systems for information sharing that have been as broadly adopted as those used in cybersecurity. There are emerging efforts to create harm taxonomies, severity criteria, and databases that could form the basis for future information sharing processes and standards. Further adoption of the AI RMF may be a catalyst for accelerating these efforts by creating requirements and demand for information as an input to the risk management process.

Recommendations:

R16. Develop a NIST Interagency or Internal Report (IR) to identify gaps in CVE and NVD with respect to cybersecurity vulnerabilities in AI systems.
R17. Engage early AI RMF adopters to identify information sharing best practices and seek input from the private sector via AISIC.

Guidance on the optimal composition of AI red teams, including different backgrounds and varying levels of skill and expertise *[RFI Sec. 1(b)]*

AI red teams should be diverse across many dimensions, including but not limited to demographics, academic training, and professional background. Diversifying AI red teams across geographic, language, and socioeconomic dimensions helps identify a broader set of risks, including risk of harms affecting marginalized communities and ways that adversaries may leverage biases or demographics to bypass protections.

Microsoft's AI red teams include people from a wide range of backgrounds, including cybersecurity, penetration testing, social engineering, national security, data science, adversarial ML, white hat hacking, cognitive neuroscience, and diversity and inclusion. AI red teams regularly leverage experts – both internal and external – from other disciplines.

Beyond the composition of an organization's AI red team, it is important that the participants in each AI red team exercise are appropriately diverse. This composition for a given AI red team exercise may vary depending on its goals; for example, an open-ended exercise may have a more varied team than an exercise focused on a specific type of harm. Part of the planning phase for each AI red team exercise should ensure that the composition of the participants is appropriate for its goals.

Microsoft approaches AI red teaming in a similar way to traditional cybersecurity red teaming, where the AI red team is independent of the developers of the AI system under test. Having an independent red team is valuable because it can provide an objective assessment of the AI system, and it can also help to identify harms that the development team may have overlooked. Other approaches, such as red teaming conducted by the product team building the AI system or by non-experts, may be appropriate for non-high risk use cases. These forms of red teaming may be less comprehensive and robust but can still effectively identify certain types of harms.

Recommendations:

R18. Supplement the *NICE Workforce Framework for Cybersecurity*[34] with Categories, Work Roles, Competencies, and Task, Knowledge, and Skill (TKS) statements related to responsible AI, including AI red teaming.

R19. Provide guidance that maps different types of risks to Competencies and TKS statements that describe AI red team exercise participants best able to identify that type of risk.

2. Reducing the Risk of Synthetic Content (E.O. 14110 Section 4.5(a))

Introduction and context

E.O. 14110 appropriately recognizes that rapid advances in AI can create new challenges and exacerbate existing risks for the information ecosystem, including the increased use of AI to generate realistic and convincing audio, video, and images that fake or alter the appearance, voice, or actions of real persons, such as political candidates, which are then misused to deceive the public (colloquially known as "deepfakes").

In addition to the risks to information and electoral integrity, AI can also be misused to generate other illegal or harmful content. For example, leading child safety organizations such as the Internet Watch Foundation have reported that AI is already being used to generate CSAM that is indistinguishable from real images, including revictimizing survivors by generating new imagery of known victims.[35] Such synthetic content is not only inherently harmful but also may be used to facilitate other harms, such as extortion, grooming, or trafficking. Large volumes of synthetic content may also hinder efforts to address

[34] *NICE Workforce Framework for Cybersecurity (https://www.nist.gov/itl/applied-cybersecurity/nice)*

[35] How AI is being abused to create child sexual abuse material (CSAM) online (https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/)

real-world harm by overwhelming law enforcement with synthetic content that is indistinguishable from real content, impeding victim identification, and fueling demands from bad actors for new content.

AI is also being misused to generate non-consensual intimate images (NCII) of real people, including high-quality synthetic content that can be used to shame, blackmail, harass, and extort adults and children alike. As early as 2019, a report by Sensity AI found that 96% of deepfakes were non-consensual intimate imagery.[36] Such synthetic content is targeting celebrities and female public figures, including as a tool for intimidation and harassment. Teen girls have also been targeted, with examples emerging where synthetic, explicit imagery has been generated and disseminated in high schools.[37] Such abuse is often highly gendered, with consequences ranging from fear and pain to lasting reputational damage.[38]

The risks from synthetic content therefore range from personal to societal. As a technology company, Microsoft has a responsibility to address potential risks arising from the abuse of our services, as well as to contribute to a safer online ecosystem, including through both safety mitigations and responsible AI design approaches. And we have well-established processes to advance digital safety and to support information integrity, including through critical partnerships and multistakeholder collaborations.[39] Many of these established frameworks and processes can be adapted to address potential risks arising from emerging technologies and, indeed, collaborative efforts are already underway to respond to and mitigate potential risks of synthetic content.

Tools, best practices, and standards for generative AI disclosure and verification

There are two common, distinct disclosure methods to help a consumer of generative AI content understand that the content was created or edited by AI: watermarking and metadata-based provenance disclosure. Microsoft's main method of disclosure for our products that use generative AI technology is to follow the Coalition for Content Provenance and Authenticity (C2PA) specification[40] to bind or attach cryptographically signed metadata at generation or post-generation. This metadata typically provides information about the history and origin of the content, such as how it was made and whether it has been edited.

Recognizing the importance of an open technical standard to enable widespread adoption of provenance by technology content creators, publishers, and platforms, Microsoft co-founded the C2PA in 2021 alongside Adobe, Arm, BBC, Intel, and Truepic. This coalition co-developed and continues to update the C2PA technical specification. Foundationally, the C2PA standard defines how a "content credential" (the disclosure data) can be cryptographically bound to common file formats, such as JPG photos or MP4 videos, in a way that is easy to display but nearly impossible to edit without being detectable. This is similar to how we might secure a physical shipping container with tamper-evident seals and then verify the contents with a verifiable bill of lading.

While C2PA is not the only solution for disclosure and verification, it is one possible disclosure standard, and it is the only provenance-based standard that exists today in a commercially viable sense. Microsoft

[36] The State of Deepfakes: Landscape, Threats, and Impact (https://medium.com/sensity/mapping-the-deepfake-landscape-27cb809e98bc)

[37] Fake Nudes of Real Students Cause an Uproar at a New Jersey High School - (https://www.wsj.com/tech/fake-nudes-of-real-students-cause-an-uproar-at-a-new-jersey-high-school-df10f1bb)

[38] The Yale Law Journal - Forum: The Continued (In)visibility of Cyber Gender Abuse (https://www.yalelawjournal.org/forum/the-continued-invisibility-of-cyber-gender-abuse?utm_source=TWITTER&utm_medium=social&utm_campaign=News+%26+Expertise)

[39] Joining the Copenhagen Pledge: a call to action for technology to empower democracy - (https://blogs.microsoft.com/on-the-issues/2022/09/22/copenhagen-pledge-cybersecurity-tech-for-democracy/)

[40] See https://c2pa.org/specifications/specifications/1.4/index.html

has chosen to adopt the C2PA specification for provenance due to several characteristics that we find to be important in a disclosure solution:

- Disclosure approaches may present privacy and human rights concerns if they convey sensitive or personally identifiable data. The C2PA specification allows for the optional omission of this data if the content creator wishes to preserve privacy.
- Malicious actors may attempt to forge disclosure information or manipulate content after it has been sealed. Using industry standard cryptographic methods, C2PA trusted signer credentials cannot be forged; thus, when content consumers see a piece of media signed, e.g., by a news publisher, political candidate, or government, they can have certainty that it indeed came from that organization. Further, if the media has been altered after it was signed, the content credential will not pass validation checks, and the consumer can detect that manipulation occurred.
- Knowing which verification tool is appropriate for which pieces of synthetic media can be challenging for content consumers and platforms alike. If a product or platform includes a verification mechanism that implements the C2PA specification, content credentials will automatically be authenticated and displayed.
- It can be challenging to convey the breath of potentially relevant information for a piece of synthetic media, including edits made across the media's lifecycle as it is picked up by new actors. The C2PA specification offers the ability to store extensive information, including optionally an edit chain history.

Watermark, fingerprint, and probabilistic generative AI detection tools are prone to errors and present reliability challenges as generative AI technology advances. However, when using verification tools aligned with the C2PA specification, there are no false positives, false negatives, or probabilities of correctness – only disclosure of the information in the content credential.

Commercial tools are bringing down the cost and reducing the difficulty of adding provenance to AI-generated or edited media, so that everyone can meaningfully disclose the role of AI in the creation and history of digital media. Microsoft has been working alongside others in C2PA to implement the specification, improving transparency and helping to drive the broader ecosystem forward. This includes the integration of provenance in Microsoft's generative AI products such as Image Creator by Designer (formerly Bing Image Creator), DALLE-3 on Azure OpenAI, and Content Credentials as a Service[41]– a new service for candidates and campaigns to add provenance to their election content.

Other methods of disclosure include watermarking, which can be visible or invisible to the naked eye. Invisible digital watermarking is a technique that embeds imperceptible pixels, sounds, or other modifications directly into the media itself, so that they can be identified later via a watermark detector or forensics. The use of digital watermarking, especially for use in synthetically generated speech, is an active area of work at Microsoft.[42]

Although we advocate for C2PA-based metadata disclosures wherever possible, in many ways, the methods of provenance and watermarking are complementary. First, they have distinct strengths: a strong use case for provenance is to determine authenticity (as provenance provides cryptographic assurance that the content credential truly came from the party signing it, and that the media has not

[41] See Content Credentials as a Service (https://blogs.microsoft.com/on-the-issues/2023/11/07/microsoft-elections-2024-ai-voting-mtac/)
[42] https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/introducing-the-watermark-algorithm-for-synthetic-voice/ba-p/3298548

been modified since last signing), while the strongest use case for watermarking is to help detect AI-generated content. Second, the two methods are complementary due to added resiliency benefits when coupled. While both provenance and watermarks can be removed by adversarial actors, state-of-the-art watermarking techniques may be more likely to withstand certain file manipulations (e.g., taking screenshots), and watermarking could be used to potentially assist with provenance metadata recovery if it is stripped. Third, provenance is amenable to an open standard for widespread disclosure and verification access while watermarking can offer additional security benefits that come with a closely held technology. Future opportunities for exploratory research on how to leverage the benefits of both technology approaches are welcomed and mentioned in the "Exploratory Approaches" section below.

Tools and best practices specific to synthetic CSAM and NCII

In addition to abuse-agnostic mitigations such as provenance disclosures and watermarking, a range of existing measures and best practices can be leveraged to prevent the specific risks of synthetic CSAM and NCII. Such mitigations must be deployed upstream to prevent generation, as part of a responsible AI infrastructure, and downstream to prevent dissemination, as part of a digital safety and content moderation infrastructure. In line with the RFI, this section focuses on the former, but a wide range of voluntary commitments[43] and regulatory guidance also provide a foundation for downstream efforts to disrupt these harms. It is also important to note that addressing both harms effectively also requires a whole-of-society response, alongside any measures by industry.

Some emerging practices follow, noting that the risk of these synthetic content harms must be considered at different layers of the AI stack, with different mitigations at each layer – each working in conjunction with the others to reduce harm to end users. Interventions will also look different between consumer and enterprise AI services:

- *Training:* Developers should be aware of the datasets used for training and take steps to avoid using any known to contain CSAM, including, where appropriate, identifying and removing any CSAM.
- *Testing and identification:* Ideally, AI models could be tested to surface specific avenues for misuse to generate synthetic CSAM, identify capabilities and limitations, and fine-tune accordingly. However, current legal frameworks create challenges that prevent robust testing for CSAM, as outlined below.
- *Mitigations:* Specific mitigation measures should be implemented and documented to address the risk of synthetic CSAM and NCII, including as a part of a responsible AI safety system, and at the application layer of the tech stack. This may include leveraging multi-modal classifiers to prevent harmful inputs as well as the output of harmful content. Hash-matching or other detection technologies could, for instance, be deployed to prevent the use of CSAM as a prompt to generate new content.
- *Other measures:* Other relevant measures include guidance and policies for users and customers, including incorporating specific prohibitions against using the AI system to generate synthetic CSAM or NCII. Developers could also prepare pre-determined responses to queries seeking CSAM or NCII.

[43] For example, the Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse (https://www.weprotect.org/library/voluntary-principles-to-counter-online-child-sexual-exploitation-and-abuse/)

Addressing these harms is complex and will require ongoing multistakeholder collaboration with AI experts.

Recommendations:

R20. Convene via AISIC industry and subject-matter experts, such as the National Center for Missing and Exploited Children (NCMEC), Internet Watch Foundation, StopNCII, and the Tech Coalition, to continue to build best practices. The Tech Coalition has, for example, already committed to support information sharing to support CSAM evaluations among its industry membership.[44]

R21. Work with industry and other experts to explore the development of best practice guidance for the systematic measurement and mitigation of CSAM risks at the model layer. Similar collaboration will be critical to ensure diverse perspectives inform sociotechnical responses to all synthetic content risks and support the development of complementary initiatives (such as educational or media literacy campaigns) by other stakeholders.

Challenges and Gaps

While considerable progress has been made to develop and deploy disclosure methods for generative AI media and to mitigate synthetic content risks, several challenges exist.

Detection: As noted above, probabilistic detection across various types of media is challenging and will likely continue to remain so. In 2019, Microsoft, together with AWS, Meta (then Facebook), the Partnership on AI, and various academics, co-sponsored a Deepfake Detection Challenge,[45] to spur researchers around the world to build new technologies that can help detect deepfakes and manipulated media. When evaluating models against a black box dataset of deepfakes, the highest-performing entrant achieved an average precision of roughly 65 percent per the results.[46] While AI detectors continue to improve, so too will AI generators get continually better at fooling AI detectors.[47]

Watermark detection: In weighing the benefits of releasing watermark detection tools, there is a careful balance between the value of keeping the watermark detector private and thus more secure, versus opening the tool to the public for more transparent verification, which allows adversaries to better learn how to evade detection. Access to a watermark detector compromises the security of the watermark, making the detector/secret key it uses best kept to limited, trusted parties rather than being public. However, while keeping watermarking approaches closely held improves how secure they are, it also complicates verification by the public of diverse media potentially leveraging a broad array of proprietary watermarks. As platforms consider how to best pass on disclosure signals to content consumers, keeping up with a growing body of watermark approaches quickly becomes challenging and authentication and display becomes infeasible without detector access.

Robustness: No disclosure method is perfect and all will be subject to adversarial attacks. This includes stripping or removal of the disclosure method and attempts to add fake disclosure signals. For instance, the potential exists for an adversary to strip a content credential from an image and then replace it with their own content credential (making an image appear to be real when it is in fact AI-generated, or vice versa). Watermarks too are subject to removal attacks through perturbations or modifications to the media by an adversary. For this reason, end-user literacy and education of what a content credential actually guarantees becomes even more important. As removal and replacement attacks will occur,

[44] Tech Coalition | Tech Coalition Hosts Generative AI Briefing for Key U.S. Stakeholders (https://www.technologycoalition.org/newsroom/tech-coalition-generative-ai-briefing)
[45] https://www.kaggle.com/c/deepfake-detection-challenge/overview
[46] https://ai.meta.com/datasets/dfdc/
[47] For more on challenges with detectors, see 2102.06109.pdf (https://arxiv.org/pdf/2102.06109.pdf)

authenticity becomes even more important, as it allows the content consumer to determine the level of confidence and trust they have in the signer of a content credential if one is displayed.

Sociotechnical challenges: While measuring technical robustness will be important, assessing societal resilience to deceptive AI content will be just as critical. Given the array of disclosure, verification, and detection methods that exist, it will be important to (1) educate the public on what labels or disclosures mean and do not mean (noting, for instance, disclosure of origin or history is neither disclosure nor verification of veracity), and (2) understand how content consumers interpret signals that media is AI-generated or edited (including, for instance, whether individuals are accurately interpreting labels or disclosures, to what extent this information is trusted, and whether trust varies depending on the actor labeling content). Similar to sociotechnical literacy around security concepts like phishing prevention, there is a need for public technology literacy campaigns focused on provenance and various disclosure methods. Similarly, for detection tools, it will be important to understand the impacts of error rates, variations in practice for false positive and false negative thresholds, and how these collectively impact the trust of users in the results detection tools convey.

Lexicon: Having a common terminology to refer to disclosure methods is a current challenge and at times, distinct methods are referred to interchangeably. While further progress is needed to align the entire ecosystem on shared terminology, the Partnership on AI's Glossary for Synthetic Media Transparency Methods[48] provides a good starting point.

Recommendation:

> **R22.** Help advance public understanding of synthetic media risks and disclosure methods and their implications through the consistent use of precise terminology. For example, NIST may consider alternatives to the term "deepfake" in official standards or guidance, as the term has no broadly accepted definition and is sometimes publicly conflated with any realistic-seeming synthetic content.

Format challenges: Methods for disclosing AI-generated some forms of audio-visual content, including images and video, are notably more mature than methods for other formats, including audio and text. We must also recognize that different formats may require different disclosure methods to work effectively. For example, provenance may not be the best solution for AI-generated text typically produced in turn-based product interfaces like chat or when used to generate sentences and code-fragments, where the resulting text is usually modified by a human and moved from one file format to another. In these scenarios, there is currently no well-defined container or unified file structure to which content credentials can be added. And while provenance and watermarks can be respectively bound to and embedded within audio files, gaps still exist for conveying these to content consumers across a variety of audio settings given user interface challenges.

Different disclosure standards and implementations: The proliferation of varying disclosure methods – such as provenance and invisible watermarks, in addition to potential future techniques – presents a significant challenge to social media platforms and other parties tasked with detecting and displaying disclosure notices. As the number of disclosure standards increases, the cost and complexity to scan for and display many indicators will also go up. This challenge is exacerbated by the burden on the public to learn and understand multiple formats to determine the authenticity or source of certain content online.

CSAM and NCII detection and classification: In addition to the detection challenges outlined above, effectively addressing synthetic CSAM and NCII may require the development of new detection

[48] https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/

capabilities. To detect and disrupt the dissemination of CSAM at scale, industry currently relies most on hash-matching technologies such as PhotoDNA. However, hash-matching only enables the identification of content that has previously been identified as harmful and hashed. It does not enable the detection of newly generated harmful content – this requires the use of classifiers and other AI and machine learning technologies. However, development of CSAM classifiers has previously been hindered because there is no way to leverage such content as training data without violating the law. It is important that policymakers consider options for a legislative safe harbor that would enable companies to leverage CSAM (including synthetic CSAM) in a limited capacity for training purposes, working with NCMEC, the Tech Coalition, and other stakeholders to develop robust, appropriate guardrails.

Red teaming and testing for CSAM: Similarly, given the current legal limitations on addressing CSAM, industry does not have a clear legal basis on which they can safely undertake the systematic testing, measurement and evaluation required to ensure AI models cannot produce synthetic CSAM. We see it as critical that policymakers work with NCMEC, industry and experts to explore options for a safe harbor enabling companies to generate the system inputs and outputs necessary to undertake robust evaluation of CSAM-related risks.

Ecosystem wide cooperation: None of these disclosure methods, mitigations, and preventive steps can be effective without ecosystem-wide adoption and multistakeholder cooperation. An invisible watermark will only be as good as the detection methods (or subsequent labeling) available to everyday consumers. Similarly, content credentials are most effective when they follow a common, open-source specification that can easily be used by generators of synthetic content and subsequently read and displayed by the media players and platforms where the content is widely viewed and distributed. To keep consumers well informed and critically engaged in their use and consumption of synthetic media content, an ecosystem wide approach is essential, where contributors to the ecosystem are held accountable to one another to maintain a cooperative, healthy information environment. Equally, to effectively address specific content risks will require cross-industry collaboration, as well as action by policymakers to deter bad actors from creating synthetic CSAM and NCII.

Exploratory approaches recommended for NIST

Microsoft looks forward to participating in AISIC and recommends NIST pursue several exploratory approaches through the Institute and with Consortium partners.

Research and measurement of maturing disclosure approaches: We believe there could be added robustness and value in potential combinations of disclosure methods for synthetic media (e.g., provenance, watermarking, and fingerprinting) to attempt to address the limitations of any one approach.

We would welcome efforts by AISIC to advance measurement methodologies to assess the performance of disclosure methods across a variety of formats. This could include expanding tests to consider emerging formats, such as interactive synthetic media, and work to advance common approaches to measure watermark durability and robustness. To address sociotechnical challenges, additional evaluation and research are needed into how disclosure technologies are being used and misused today, and the extent to which these disclosures impact the trust of the general public.

Recommendations:

> **R23.** Develop evaluation infrastructure to study how provenance and watermarking technologies are being used, to what extent the technologies are understood and trusted when they are used as

intended and when they are misused/abused, and how trust in labels varies depending on the entity labeling content as AI-generated or edited.

R24. Explore, in conducting technical assessments, the impact and benefits of combining disclosure methods (e.g., provenance, watermarking, and/or fingerprinting), including in the face of adversarial attacks.

Exploratory, proof of concept research: AISIC-driven research would also be beneficial to seek to address current disclosure gaps.

Recommendations:

R25. Consider, including through AISIC, opportunities to facilitate the mobilization of research breakthroughs on disclosure methods for AI-generated text, with a focus on methods that are robust to attacks and work for short-form text.

R26. Consider driving research on whether a standardized approach to watermarking might be feasible for audio-visual media given the public/private key challenges noted and the potential that may exist for a hybrid approach.

Education based on AISIC findings: We encourage NIST, in collaboration with other agencies, to leverage AISIC findings to foster media literacy, so that citizens learn about both the risks of synthetic content and tools available to better protect themselves from being manipulated or deceived when such content is misused. This would help ensure citizens are critical content consumers and help ensure that as provenance and other complementary disclosure methods are deployed at scale, they are easily digested and comprehended.

Auditing

Microsoft would also welcome contributions by NIST to create pre-standardization materials to enable assessment and verification of tools used to detect AI-generated content and verify provenance.

Recommendation:

R27. Create pre-standardization materials addressing the performance of probabilistic detection tools that detect generative AI content as well as the verification and authentication tools used to view and analyze provenance metadata. Work on the latter should address what file formats must be readable, viewable, and displayed for both industry cooperation and public accessibility**.**

3. Advance responsible global technical standards for AI development (E.O. 14110 Section 11(b))

NIST's global engagement to promote and develop consensus-based standards for AI should align with the work of the American National Standards Institute (ANSI),[49] the United States Standards Strategy (USSS),[50] and the principles established in the AI RMF.[51]

NIST should consult and collaborate with civil society, academia, and industry stakeholders on NIST's AI work through the model of a public-private partnership to determine when and how standardization is best applied to AI. NIST should further support these efforts through constructive engagement with international partners and other relevant standards development organizations (SDOs). Such collaboration will support the development and maintenance of both high-level and application-specific

[49] https://www.ansi.org/

[50] https://share.ansi.org/Shared%20Documents/Standards%20Activities/NSSC/USSS-2020/USSS-2020-Edition.pdf

[51] https://www.nist.gov/itl/ai-risk-management-framework

AI standards. NIST should contribute both as a stakeholder and as a neutral convenor of open processes to establish pre-standardization materials for AI.

The topic areas that should be reflected and maintained in high-level AI standards include those that would be "horizontal" in nature and therefore applicable across virtually all types of AI systems, which makes them particularly impactful, for example:

- Foundational elements, such as AI terminology, use cases, and reference architectures.
- Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis.
- Measurement and evaluation of inclusivity, fairness, accountability, and representativeness (i.e., non-discrimination, representation of lower resourced languages, and the need for data to reflect freedom of expression) in the collection and use of data, guidelines, and standards for trustworthiness, verification, and assurance of AI systems.

By engaging in the identification and/or development of these high-level standards, NIST should consider existing specifications to avoid creating unnecessary duplication, which creates market confusion. For example, as new terminology and nomenclature are developed for AI, NIST should consider contributing to the revision of existing standards rather than creating novel specifications that may contradict or conflict with existing work. In addition, when developing standards related to the trustworthiness of AI systems, NIST should evaluate the work of multiple standardization venues, including a broad range of applicable standardization bodies, consortia groups, and open source software activities. This will improve international coherence and more effectively address interoperability-related issues, which in turn benefits all stakeholders in the global marketplace. As previously referenced in this response, the specification related to provenance by technology content creators, publishers, and platforms that has been done by C2PA[40] is a good example of consensus-based work that is being used today to advance responsible disclosure practices.

NIST should engage with, and promote, standards that enable 3rd party assessment of management practices that help organizations demonstrate conformity to AI rules and regulations. An example of this is the AI Management System (AIMS) (ISO/IEC 42001). Microsoft recognizes that diverse regulatory and governance frameworks will be promulgated over time. The advantage of AIMS is that it facilitates developers and implementers of AI systems' ability to address the implications of both risks and harms to individuals, organizations, and society. This allows the providers of AI-based solutions to both prioritize the rights of individuals over organizational risk while innovating to the benefit of all stakeholders. Balancing risks and harms begins with common terminology and Microsoft agrees with the NIST AI Risk Management Framework approach. From there, Microsoft assesses the capabilities and risks of an AI system as well as the likelihood of harms occurring from those risks. We believe it is the combination of risk management frameworks and impact assessments that facilitates a comprehensive approach for organizations building or deploying AI systems in a global context.

The last category of standards to consider are application AI standards which cover either a particular technology or sector-specific uses of AI~~.~~ Examples of application-specific standards include:

- Topologies of AI systems.
- Human-computer interface design for AI systems.
- Application-specific standards more broadly (e.g., for computer vision or facial recognition technology).

NIST should seek to leverage concepts from high-level standards and apply them to application-specific standards for consistency across all standards types.

In another recent NIST RFI for the U.S. National Standards Strategy for Critical and Emerging Technology, Microsoft submitted comments that are relevant here.[52] NIST should coordinate and engage in AI international standardization efforts with the following principles and concepts guiding its related decisions and actions:

- Value a diversity of standardization bodies including industry consortia as well as formal international organizations.
- Avoid government interests picking (or being perceived to pick) standards "winners and losers".
- Standards are intended to promote interoperability, competition, and overall market growth/coherence, and should not be written to require uniformity among competing products and services.
- Standardization and open source software are complementary and not mutually exclusive.

Microsoft encourages NIST to recognize that there are thousands of relevant open source projects that are AI-related in the marketplace as well as innovations taking place in private labs. We suggest that NIST examine some of the many scenarios where technology-focused consensus-building can effectively leverage open source projects and outcomes in conjunction with standardization efforts and other deliverables. For example, there are no broadly adopted technical standards for the training of AI models, rather there are academic papers and open source projects that have coalesced into common practice. The lack of a documented standard does not invalidate the collaboration or learnings across the AI development community. Situations like this show the importance of leveraging existing collaborations as both pre-standardization activity and a potential contribution to new standards development.

AISIC represents an opportunity for NIST to convene a broad spectrum of stakeholders to produce a mix of technical and management practice outcomes. We are not advocating that every action of AISIC should result in a standard. Instead, NIST should look to the outputs from AISIC to:

- Engage in pre-standardization efforts across a combination of methods for collaboration to determine when and how standards are best used. The overlay of governance practices, conformity assessment, security practices, etc. may be based on standards even if the underlying technical practice is through other mechanisms.
- Be an active stakeholder in the AI-related standardization process. This includes participating and providing technical contributions in a broad range of AI-related standardization activities and representing and promoting the needs of the U.S. government as a consumer and user of AI-related technologies.
- Contribute standards-related proposals to related international standardization activities; and
- Leverage international standards whenever possible (rather than developing related national standards) as this will lead by example and help achieve coherence that will benefit all stakeholders.

[52] https://www.regulations.gov/comment/NIST-2023-0005-0061